RICE UNIVERSITY

**Evolutionary Genetics of Dictyostelids: Cryptic Species, Sociality and Sex**

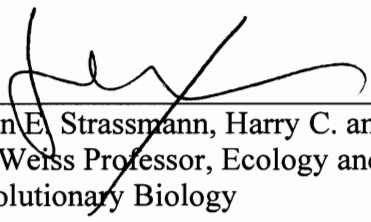
by

**Sara Edith Kalla**

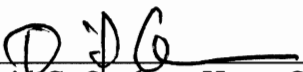
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS OF THE DEGREE

**Doctor of Philosophy**

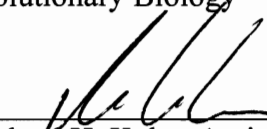
APPROVED, THESIS COMMITTEE:



Joan E. Strassmann, Harry C. and Olga K. Weiss Professor, Ecology and Evolutionary Biology



David C. Queller, Harry C. and Olga K. Weiss Professor, Ecology and Evolutionary Biology



Michael H. Kohn, Assistant Professor, Ecology and Evolutionary Biology

Luay K. Nakhleh, Associate Professor, Computer Science

HOUSTON, TEXAS
MARCH 2011

**Abstract**

**Evolutionary Genetics of Dictyostelids:**

**Cryptic Species, Sociality and Sex**

By

Sara Edith Kalla

*Dictyostelium discoideum* serves as an ideal system to study social evolution because of the social stage of its lifecycle, where individuals aggregate to build a multicellular structure. However, much of its basic biology remains unknown and this limits its utility. I used three separate projects to fill these gaps.

In my first project, I examined how speciation and genetic diversity affects kin discrimination using a related dictyostelid, *Polysphondylium violaceum*. I sequenced the ribosomal DNA of 90 clones of *P. violaceum* and found that *P. violaceum* is split into several morphologically identical groups. When allowed to cooperate in pairwise mixes, I found that some clones cooperated with others in their group, but in mixes between groups, clones did not cooperate.

For my second project, I looked at whether *D. discoideum* has sex in natural populations. While sex has been observed in laboratory clones of *D. discoideum*, it is unclear whether sex occurs in natural populations, and sex can influence the evolution of traits. I used a dataset of

microsatellites in 24 clones of *D. discoideum* to look for a decrease in linkage disequilibrium as a molecular sign of sex. Linkage disequilibrium is higher between physically close loci than between loci on different chromosomes. From this, I conclude that *D. discoideum* undergoes recombination in nature.

Lastly, I used the genome sequence of *D. discoideum* to look at large scale patterns of evolution. Mutations tend to be biased towards A/T from G/C so, on average, mutations should lower the nucleotide content of sequences. The removal of these mutations, purifying selection, should preserve nucleotide content. I used the genomes of *D. discoideum* and *Plasmodium falciparum* identify classes of sequences that should be under different amounts of purifying selection and compared their nucleotide contents. In all cases, those sequences under more purifying selection had higher GC contents than sequences under less purifying selection. Looking at relative nucleotide content may thus serve as an indicator purifying selection.

These three studies add insight on how cooperation works in dictyostelids as well as adding an understanding of how traits, social and otherwise, would evolve in this system.

**Acknowledgements**

I would like to thank my advisors Dave Queller and Joan Strassmann for all their help, encouragement and guidance. I would also like to thank my committee members Luay Nakhleh and Michael Kohn for their insights and patience. Much appreciation goes to my benchmate, Jennie Kuzdzal-Fick and my officemate, Chandra Jack for many productive conversations. Beyond the hedges of Rice, I would like to thank my Keck Fellows group for encouragement in the realm of computational biology. I would also like to thank my parents for never discouraging questions and showing me where to find answers. Last but not least, I would like to thank PJ Holmes for sanity and support and Freyja for much needed balance.

**Table of Contents**

**List of Tables**

**List of Figures**

# CHAPTER 1

# Introduction

In this dissertation, I cover three projects that contribute to the knowledge of the biology of the social amoebae. First, I look at kin discrimination and phylogeny in *Polysphondylium violaceum,*a species in the Dictyostelidae. Second, I examine linkage disequilibrium in *Dictyostelium discoideum*, and evaluate methods for doing this generally. Third, I examine the effects of selection pressure on nucleotide content in *D. discoideum* and another GC poor organism, *Plasmodium falciparum*.

*D. discoideum* is one of 13 model organisms recognized by the NIH (http://www.nih.gov/science/models). Traditionally, *D. discoideum* has been used to look at multicellular development because individual amoebae aggregate to form a multicellular structure during the social stage of its lifecycle (Devreotes 1989; Kessin 2001). The organism has also recently been used to look at the evolution of social interactions (Strassmann, Zhu, and Queller 2000; Kessin 2001). Despite this status, much of the basic biology of the dictyostelids is not well understood. For example, an understanding of the population genetics can contribute to a greater understanding of how sociality is maintained. In sexual populations, different individuals will have varying degrees of relatedness while asexual populations will be clonal, with some individuals being identical. This population structure can influence the evolution of social traits, or ones that rely on interactions with other individuals. To fully use

*D. discoideum* as a model organism, an understanding of the basic biology of the species is necessary.

*D. discoideum* was formally described in 1935 (Raper 1935), but since the development of axenic clones in the 1960s (Watts and Ashworth 1970), most of the work in this system has focused on axenic clones developed in the laboratory from a single wild isolate. To develop this system as a model system for sociality, additional work is necessary. Studies on social behaviors in other dictyostelids can give us insight to the evolution of social behavior in this system. Studies on sex and speciation give us information on the relationships between isolates as well as how recognition genes spread in populations. Previous studies of dictyostelids have shown differing behaviors in social situations, with some species discriminating in favor of kin, and some species where individual strains cooperate with others (Kaushik, Katoch, and Nanjundiah 2006; Mehdiabadi et al. 2006; Ostrowski et al. 2008).

To trace the evolution of kin discrimination in this group, I chose to study *Polysphondylium violaceum*, which is the basal member of the group that the well-studied *D. discoideum* belongs to. Study of the biology of *P. violaceum* is complicated by the presence of at least two morphologically identical species. I found that *P. violaceum* is likely composed of two or more genetically separate species, and further,

individual clones of these species do not cooperate with individuals of the other species in the social stage of their lifecycle.

Through laboratory studies, we know that *D. discoideum* can have sex, but until recently, whether they do so in nature was unknown (Flowers et al. 2010). Because identifying matings in the wild is not feasible, a molecular approach of looking for signs of recombination is necessary. Linkage disequilibrium arises when alleles at one locus are preferentially associated with the alleles at another locus. Recombination decreases linkage disequilibrium by changing the associations between the two loci. Using single nucleotide polymorphisms (SNPs), Flowers et al. (2010) found that *D. discoideum* does undergo recombination. I examined the levels of linkage between 100 microsatellite loci located throughout the genome, comparing the levels of linkage disequilibrium of loci that are close together with those that are farther apart or on different chromosomes. Studies have avoided using microsatellites when looking at linkage disequilibrium because of biases inherent in microsatellite data (Ardlie, Kruglyak, and Seielstad 2002). However, resampling has been proposed as a correction measure (Devlin et al. 2001). Here I propose a novel correction metric based on information about average linkage disequilibrium and evaluate both correction measures.

The completion of the genome sequence of *D. discoideum* presents an opportunity to examine how non-adaptive forces, such as mutational bias, and selection combine to influence the GC content of a sequence. *D. discoideum* is exceedingly GC poor, with an average GC content of roughly 19.4% (Eichinger et al. 2005). This suggests that the GC content of the genome of *D. discoideum* is dominated by a mutational bias towards AT from GC. Purifying selection should remove these mutations, preserving a higher GC content. I split the genome of *D. discoideum* into paired sets of sequences based on the relative amounts of purifying selection that sequence had, such as coding regions and non-coding regions, then compared the GC contents of these different classes of sequences.

These molecular projects are of further interest because they are not just applicable to the cellular slime molds. Linkage disequilibrium is widely used in a variety of applications, most notably association mapping. Biases of linkage disequilibrium measures and background linkage disequilibrium prevent effective identification disease markers(Ardlie, Kruglyak, and Seielstad 2002). Using the most accurate correction measure can improve this identification. Lewontin's D' is a frequently used measure of linkage disequilibrium because it can be used with multiallelic data such as microsatellites, and because it is standardized so that it

ranges from 0 to 1 (Lewontin 1988). However, D' is still biased by sample size and allele frequency (Ardlie, Kruglyak, and Seielstad 2002; McRae et al. 2002). I also compare several different correction methods including using resampling (Devlin et al. 2001) and using D' between different chromosomes to normalize D'. An understanding of the effects of purifying selection on nucleotide content can help disentangle the effects o f purifying selection from other selection. It also helps bridge the divide between the theoretical understanding of how selection pressures change nucleotide sequences and the molecular mechanisms by which those sequences change.

These three projects contribute to our understanding of the social amoeba. Looking at kin discrimination in *P. violaceum* helps add to our knowledge of the evolutionary context of this behavior in the Dictyostelids. Studies on linkage disequilibrium and the interaction of mutation and purifying selection help us understand the mechanisms of evolution in this system.

**References:**

Ardlie, K. G., L. Kruglyak, and M. Seielstad. 2002. Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics **3**:299-309.

Devlin, B., K. Roeder, C. Otto, S. Tiobech, and W. Byerley. 2001. Genome-wide distribution of linkage disequilibrium in the population of Palau and its implications for gene flow in Remote Oceania. Human Genetics **108**:521-528.

Devreotes, P. 1989. Dictyostelium discoideum - a model system of cell-cell interactions in development. Science **245**:1054-1058.

Eichinger, L., J. A. Pachebat, G. Glockner, M. A. Rajandream, R. Sucgang, M. Berriman, J. Song, R. Olsen, K. Szafranski, Q. Xu, B. Tunggal, S. Kummerfeld, M. Madera, B. A. Konfortov, F. Rivero, A. T. Bankier, R. Lehmann, N. Hamlin, R. Davies, P. Gaudet, P. Fey, K. Pilcher, G. Chen, D. Saunders, E. Sodergren, P. Davis, A. Kerhornou, X. Nie, N. Hall, C. Anjard, L. Hemphill, N. Bason, P. Farbrother, B. Desany, E. Just, T. Morio, R. Rost, C. Churcher, J. Cooper, S. Haydock, N. van Driessche, A. Cronin, I. Goodhead, D. Muzny, T. Mourier, A. Pain, M. Lu, D. Harper, R. Lindsay, H. Hauser, K. James, M. Quiles, M. M. Babu, T. Saito, C. Buchrieser, A. Wardroper, M. Felder, M. Thangavelu, D. Johnson, A. Knights, H. Loulseged, K. Mungall, K. Oliver, C. Price, M. A. Quail, H. Urushihara, J. Hernandez, E. Rabbinowitsch, D. Steffen, M. Sanders, J. Ma, Y. Kohara, S. Sharp, M. Simmonds, S. Spiegler, A. Tivey, S. Sugano, B. White, D. Walker, J. Woodward, T. Winckler,

Y. Tanaka, G. Shaulsky, M. Schleicher, G. Weinstock, A. Rosenthal, E. C. Cox, R. L. Chisholm, R. Gibbs, W. F. Loomis, M. Platzer, R. R. Kay, J. Williams, P. H. Dear, A. A. Noegel, B. Barrell, and A. Kuspa. 2005. The genome of the social amoeba Dictyostelium discoideum. Nature **435**:43-57.

Flowers, J., S. Li, A. Stathos, G. Saxer, E. Ostrowski, D. Queller, J. Strassmann, and M. Purugganan. 2010. Variation, sex, and social cooperation: molecular population genetics of the social amoeba *Dictyostelium discoideum*. Plos Genetics **6**.

Kaushik, S., B. Katoch, and V. Nanjundiah. 2006. Social behaviour in genetically heterogeneous groups of *Dictyostelium giganteum*. Behavioral Ecology and Sociobiology **59**:521-530.

Kessin, R. H. 2001. Dictyostelium: evolution, cell biology, and the development of multicellularity. Developmental and Cell Biology Series **38**:i-xiv, 1-294.

Lewontin, R. C. 1988. On measures of gametic disequilibrium. Genetics **120**:849-852.

McRae, A. F., J. C. McEwan, K. G. Dodds, T. Wilson, A. M. Crawford, and J. Slate. 2002. Linkage disequilibrium in domestic sheep. Genetics **160**:1113-1122.

Mehdiabadi, N. J., C. N. Jack, T. T. Farnham, T. G. Platt, S. E. Kalla, G. Shaulsky, D. C. Queller, and J. E. Strassmann. 2006. Kin preference in a social microbe - Given the right circumstances, even an amoeba chooses to be altruistic towards its relatives. Nature **442**:881-882.

Ostrowski, E. A., M. Katoh, G. Shaulsky, D. C. Queller, and J. E. Strassmann. 2008. Kin Discrimination Increases with Genetic Distance in a Social Amoeba. Plos Biology **6**:2376-2382.

Raper, K. B. 1935. Dictyostelium discoideum, a new species of slime mold from decaying forest leaves. Journal of Agricultural Research **50**:0135-0147.

Strassmann, J. E., Y. Zhu, and D. C. Queller. 2000. Altruism and social cheating in the social amoeba Dictyostelium discoideum. Nature **408**:965-967.

Watts, D. J., and J. M. Ashworth. 1970. Growth of myxamoebae of cellular slime mould Dictyostelium discoideum in axenic culture. Biochemical Journal **119**:171-174.

# CHAPTER 2

## Kin discrimination and possible cryptic species in the social amoeba *Polysphondylium violaceum*.

**Abstract:**

The genetic diversity of many protists is unknown. The differences that result from this diversity can be important in interactions among individuals. The social amoeba *Polysphondylium violaceum*, which is a member of the Dictyostelia, has a social stage where individual amoebae aggregate together to form a multicellular fruiting body with dead stalk cells and live spores. Individuals can either cooperate with amoebae from the same clone, or sort to form clonal fruiting bodies. In this study we look at genetic diversity in *P. violaceum* and at how this diversity impacts social behavior. The phylogeny of the ribosomal DNA sequence (17S to 5.8S region) shows that *P. violaceum* is made up of at least two groups. Mating compatibility is more common between clones from the same phylogenetic group, though matings between clones from different phylogenetic groups sometimes occurred. *P. violaceum* clones are more likely to form clonal fruiting bodies when they are mixed with clones from a different group than when they are mixed with a clone of the same group. Both the phylogenetic and mating analyses suggest the possibility of cryptic species in *P. violaceum.* The level of divergence found within *P. violaceum* is comparable to the divergence between sibling species in other dictyostelids. Both major groups A/B and C/D/E/F show kin discrimination, which elevates relatedness within fruiting bodies but not to

the level of clonality. The diminished cooperation in mixes between groups suggests that the level of genetic variation between individuals influences the extent of their cooperation.

**Introduction**:

Identifying cryptic species is important; morphological similarity may mask great differences in physiology, ecology, and behavior (Saez and Lozano 2005). For example, *Oreaster reticulatus* starfish preferentially prey on only one of two sympatric cryptic species of Caribbean fire sponges (*Tedania ignis* and *T. klausi*) (Wulff 2006). Sympatric cryptic species of African weakly electric fishes (*Campylomormyrus tamandua* and *C. numenius*) exhibit different patterns of electric organ discharge that these fishes use for both electrolocation and communication (Feulner et al. 2006). In these cases, identifying the species has led to a greater understanding of the variation in these traits. In African weakly electric fishes, this variation in communication affects interactions between individuals such as mate recognition and mate choice.

Social behavior can be doubly impacted by cryptic speciation. In addition to differences in behavior between the two species, social interactions are dependent on the relationship between the interactors. If the interactors come from different species, then the individuals should be

much less likely to perform altruistic acts. For example, in the two parapatric wood ant cryptic species *Formica lugubris* and *F. paralugubris*, workers exhibit discrimination against brood from the other sibling species when the workers are returning exposed brood to the nest (Maeder, Freitag, and Cherix 2005). However, the two species do not always discriminate against brood that is from the same species but from a different nest (Maeder, Freitag, and Cherix 2005).

Dictyostelids, or social amoebae, have a complex life cycle that includes social behavior and altruism at a certain stage in their life history (Figure 2.1). They are unicellular haploid eukaryotes that live in soil and consume bacteria. When their food source is depleted, they aggregate together into a mound of cells, which then proceeds along one of two different forms of development. In the sexual cycle, two cells of compatible mating types fuse to form a giant cell where the nuclei fuse and undergo meiotic recombination (Clark, Francis, and Eisenberg 1973; Nickerson and Raper 1973a) (Figure 2.1). The giant cell engulfs surrounding cells and eventually encysts. In nature haploid, recombined daughter cells eventually hatch from the cysts, though this is not easily achieved in the laboratory (Flowers et al. 2010). In the social stage, the aggregation organizes into one or more multicellular slugs. These slugs then develop into fruiting bodies. During fruiting body formation, some of

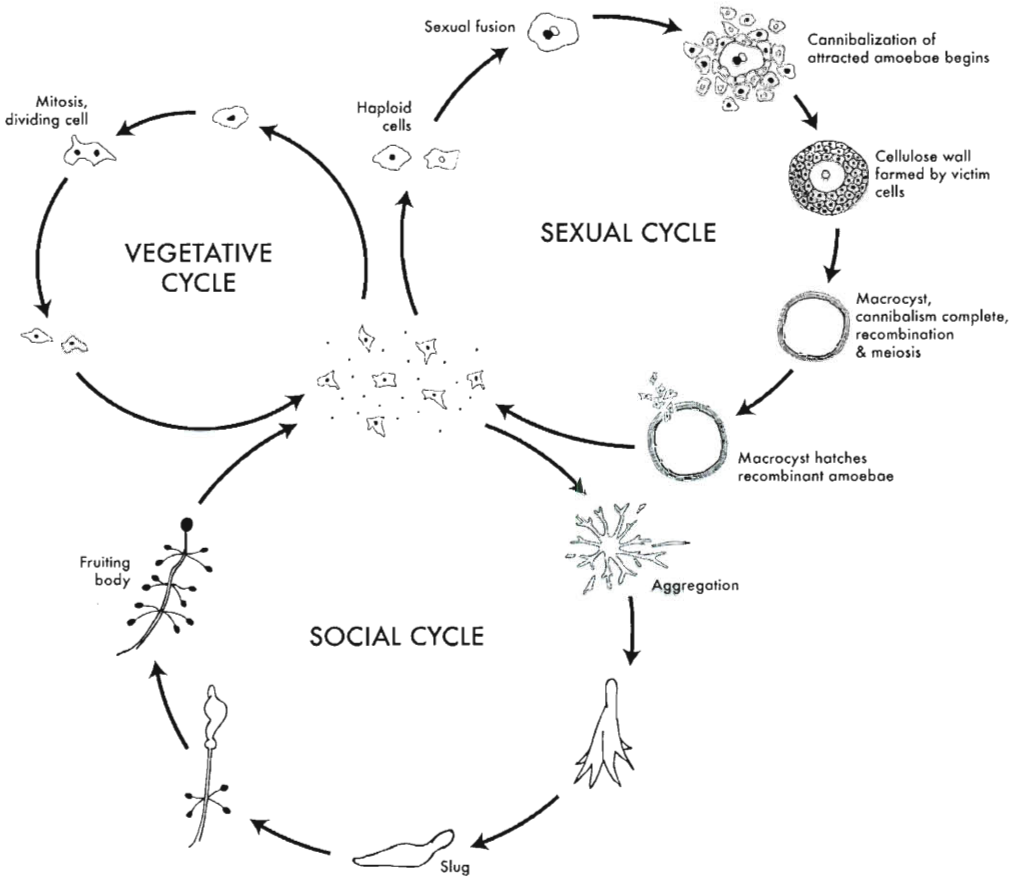



**Figure 1.1**. Life cycle of *Polysphondylium violaceum*. Most of its life, this haploid social amoeba undergoes the vegetative cycle, preying upon bacteria in the soil, and periodically dividing mitotically. When food is scarce, either the sexual cycle or the social cycle begins. Under the social cycle, amoebae aggregate to glorin by the thousands, and form a motile slug, which moves towards light. Ultimately the slug forms a fruiting body in which some of the cells die to lift the remaining cells up to a better place for sporulation and dispersal. Under the sexual cycle, amoebae aggregate to glorin and sex pheromones, and two cells of opposite mating types fuse, and then begin consuming the other attracted cells. Before they are consumed, some of the prey cells form a cellulose wall around the entire group. When cannibalism is complete, the giant diploid cell is a hardy macrocyst, which eventually undergoes recombination and meiosis, and hatches hundreds of recombinants. Not drawn to scale. Image credit: David Brown and Joan Strassmann, under Creative Commons Attribution Share-Alike 3.0 license.

the cells die to form a stalk and other cells form hardy spores at the top of the stalk. Because stalk cells die, they should be expected to preferentially form fruiting bodies with identical or closely related clones and discriminate against individuals and non-kin by sorting from them and forming independent, clonal, fruiting bodies.

There have been several studies of the diversity of individual species of dictyostelids (Clark, Francis, and Eisenberg 1973; Clark 1974; Briscoe et al. 1987; Mehdiabadi et al. 2009; Mehdiabadi et al. 2010), but overall there has been little study on the diversity within any one species. It has been suggested that *Polysphondylium violaceum* is a cryptic species complex composed of at least two separate morphologically identical species (Clark, Francis, and Eisenberg 1973; Clark 1974). Clark examined 49 clones of *P. violaceum* collected in Massachusetts for macrocyst formation. Clark observed two different groups within *P. violaceum* that each formed macrocysts when individuals from the same group were paired (Clark 1974). She did not further characterize the two putative species, and the clones are not available for further study.

Kin discrimination has been observed in *D. discoideum, D. giganteum,* and *D. purpureum* (Kaushik, Katoch, and Nanjundiah 2006; Mehdiabadi et al. 2006; Ostrowski et al. 2008), but no such work has been done with *P. violaceum*. It differs from the previously studied species in its

branching fruiting body with many small clumps of spores (Raper 1984), and it is basal to group 4 dictyostelids (Schaap et al. 2006). It is not phylogenetically close to most other species that were also called *Polysphondylium* because they were classified on the basis of their branched fruiting bodies, and not phylogenetics (Mediabadi et al. 2006).

We examined cryptic speciation and kin recognition in *P. violaceum*. We used both DNA sequence data and mating data to look at both the population structure of *P. violaceum* and how this structure relates to cooperation in the social stage. We sequenced ribosomal DNA of 90 clones of *P. violaceum* and constructed a gene tree to examine population structure. We also performed mating experiments to understand patterns of potential gene flow. We tested for cooperation and discrimination by performing 13 mixes of cells from pairs of clones, that were then allowed to develop into the social stage, so sorting could be investigated.

**Methods:**

A. Collection of clones

We collected 80 clones from undisturbed areas of the Houston Arboretum and Nature Center, Houston, TX (27 clones); Brazos Bend State Park, Needville, TX (1 clone); Mountain Lake Biological Station,

Mountain Lake, VA (16 clones); Linville Falls, NC (1 clone); Urbana, IL (22 clones); and Heidelberg, Germany (9 clones) (see Table 2.1 for details). *P. violaceum* is a cosmopolitan species, found throughout the world including the Americas (Sutherland and Raper 1978; Vadell and Cavender 1998; Vadell 2000), Europe (Cavender, Cavender-Bares, and Hohl 1995; Romeralo and Lado 2006), Asia (Hagiwara 1990; Hagiwara 2000; Yeh 2003), and Australia (Landolt et al. 2008).

To ensure that each sample had only one genotype, we clonally isolated *P. violaceum* from soil samples. We cultured them on hay infused agar plates (1 L hay infused $H_2O$ (15 g hay left in 1.5 L $H_2O$ overnight, then filtered), 1.5 g $KH_2PO_4$, 0.62 g $Na_2HPO_4$, 15 g agar, autoclaved) with *Klebsiella aerogenes* as a food source. We then replated them so that individual cells grew into discrete colonies. We harvested one colony from each sample to ensure the clones were clonal. Table 2.1 has a complete list of the clones. We obtained five clones from South Africa (RSA clones) from J. Landolt and acquired the clone P6 and four clones from Wisconsin (WS clones) from the Dictybase stock center (P6 depositor: P. Schaap, WS clones depositor: G. Erdos, (Fey et al. 2006)). Initially, we used the morphology of fruiting bodies to identify clones as *Polysphondylium violaceum*. *P. violaceum* has a unique fruiting body structure, with each stalk supporting multiple whorls of spore containing sori (3-5 sori per

| Haplotype | ID | Clone | Location | GPS |
|---|---|---|---|---|
| 1 | QSvi1 | 3B1-16 | Texas- Brazos Bend | 29 23 38.08 N, 98 45 11.88 W |
| | QSvi12 | H11A4 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi14 | H11B2 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 10 | QSvi24 | H22A4 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi57 | PV25A2C | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi77 | V7E2 | Virginia? | 37 22 21.62 N 80 34 09.76 W |
| | QSvi13 | H11B1 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi15 | H12A1 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi17 | H17B1 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi20 | H20B4 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi23 | H22A3 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi26 | H22B5 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 11 | QSvi27 | H22B6 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi28 | H26B1 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi29 | H26B2 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi30 | H26B3 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi56 | PV25A2 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi80 | V7I1 | Virginia? | 37 22 21.62 N 80 34 09.76 W |
| | WS602 | WS602 | Wisconsin | unknown |
| 12 | QSvi16 | H14B1 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| | QSvi62 | TV23B | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 13 | QSvi18 | H20B1 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |

**Table 2.1** *Polysphondylium violaceum* clones used in this study

| | | | | |
|---|---|---|---|---|
| 14 | QSvi19 | H20B3 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 15 | QSvi21 | H21A1 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 16 | QSvi22 | H22A2 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 17 | QSvi25 | H22B4 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 18 | QSvi31 | H4B1 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 19 | QSvi32 | IL10C | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| | QSvi33 | IL10D | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 2 | QSvi2 | GID2 | Heidelberg, Germany | 49 23 12.13 N, 8 42 34.70 E |
| | QSvi4 | GIIA1P | Heidelberg, Germany | 49 23 12.13 N, 8 42 34.70 E |
| 20 | QSvi34 | IL11B | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 21 | QSvi35 | IL12A | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 22 | QSvi36 | IL14A | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 23 | QSvi37 | IL14B | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 24 | QSvi38 | IL15A | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 25 | Qsvi39 | IL15B | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 26 | QSvi40 | IL16B | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 27 | QSvi41 | IL17B | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| | QSvi49 | IL6A | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 28 | QSvi42 | IL17C | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 29 | QSvi43 | IL1A | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 3 | QSvi3 | GIE | Heidelberg, Germany | 49 23 12.13 N, 8 42 34.70 E |
| 30 | QSvi44 | IL1B | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |

**Table 2.1 cont.** *Polysphondylium violaceum* clones used in this study

| | | | | |
|---|---|---|---|---|
| 31 | QSvi45 | IL21D | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 32 | QSvi46 | IL22D | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 33 | QSvi47 | IL3D | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 34 | QSvi48 | IL5D | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 35 | QSvi50 | IL6B | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 36 | QSvi51 | IL6C | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 37 | QSvi52 | IL9B | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 38 | QSvi53 | IL9D | Urbana-Champaign, IL | 40 06 21.63 N, 88 13 37.26 W |
| 39 | QSvi54 | NC29K1 | North Carolina- Linville Falls | 35 57 36.46 N, 81 56 32.78 W |
| 4 | QSvi5 | GVI3 | Heidelberg, Germany | 49 23 12.13 N, 8 42 34.70 E |
| | QSvi10 | GVV5 | Heidelberg, Germany | 49 23 12.13 N, 8 42 34.70 E |
| 40 | P6 | P6 | unknown | unknown |
| 41 | QSvi55 | PV25A1A | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 42 | QSvi58 | PVS-15 | South Africa | 29 00 00 S, 24 00 00 E |
| 43 | RSA21B | RSA21B | South Africa | 29 00 00 S, 24 00 00 E |
| | RSA22B | RSA22B | South Africa | 29 00 00 S, 24 00 00 E |
| 44 | RSA23B | RSA23B | South Africa | 29 00 00 S, 24 00 00 E |
| 45 | RSA9B | RSA9B | South Africa | 29 00 00 S, 24 00 00 E |
| 46 | RSA19A | RSA19A | South Africa | 29 00 00 S, 24 00 00 E |
| 47 | QSvi59 | TV15C | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 48 | QSvi60 | TV19B | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 49 | QSvi61 | TV21C | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |

**Table 2.1 cont.** *Polysphondylium violaceum* clones used in this study

| | | | | |
|---|---|---|---|---|
| 5 | QSvi6 | GVII2 | Heidelberg, Germany | 49 23 12.13 N, 8 42 34.70 E |
| 50 | QSvi63 | TV6A | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 51 | QSvi64 | TV8B | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 52 | QSvi65 | TV1A | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |
| 53 | QSvi66 | V313B1 | Virginia MLBS | 37 22 21.62 N 80 34 09.76 W |
| 54 | QSvi67 | V313C | Virginia MLBS | 37 22 21.62 N 80 34 09.76 W |
| | QSvi68 | V319D3 | Virginia MLBS | 37 22 21.62 N 80 34 09.76 W |
| 55 | QSvi69 | V321B1 | Virginia MLBS | 37 22 21.62 N 80 34 09.76 W |
| | QSvi70 | V322C1 | Virginia MLBS | 37 22 21.62 N 80 34 09.76 W |
| | QSvi71 | V342A3 | Virginia MLBS | 37 22 21.62 N 80 34 09.76 W |
| 56 | QSvi72 | V621D1 | Virginia MLBS | 37 22 21.62 N 80 34 09.76 W |
| | QSvi73 | V621D2 | Virginia MLBS | 37 22 21.62 N 80 34 09.76 W |
| 57 | QSvi74 | V632B1 | Virginia MLBS | 37 22 21.62 N 80 34 09.76 W |
| 58 | QSvi75 | V7C1 | Virginia? | 37 22 21.62 N 80 34 09.76 W |
| | QSvi79 | V7G1 | Virginia? | 37 22 21.62 N 80 34 09.76 W |
| 59 | QSvi76 | V7D1 | Virginia? | 37 22 21.62 N 80 34 09.76 W |
| 6 | QSvi7 | GVII9 | Heidelberg, Germany | 49 23 12.13 N, 8 42 34.70 E |
| 60 | QSvi81 | V80A1 | Virginia MLBS | 37 22 21.62 N 80 34 09.76 W |
| 61 | WS598 | WS598 | Wisconsin | unknown |
| 62 | WS600 | WS600 | Wisconsin | unknown |
| 63 | WS601 | WS601 | Wisconsin | unknown |
| 7 | QSvi8 | GVIII2 | Heidelberg, Germany | 49 23 12.13 N, 8 42 34.70 E |
| 8 | QSvi9 | GVIII7 | Heidelberg, Germany | 49 23 12.13 N, 8 42 34.70 E |
| 9 | QSvi11 | H11A1 | Houston Arboretum | 30 02 04.10 N, 95 23 14.32 W |

**Table 2.1 cont.** *Polysphondylium violaceum* clones used in this study

whorl) at regularly spaced intervals and a solitary sori at the end. The sori range in color from lavender to violet (for a complete description, see (Raper 1984)).

B. Genetic analysis

To look at the relationships between wild clones of *P. violaceum*, we sequenced a ~2500 bp region that included the 17S, internal transcribed spacer 1, and 5.8S RNA (17S-5.8S) of each clone. The 17S rDNA sequence has already been used to look at the phylogeny of the entire group of dictyostelids (Schaap et al. 2006; Romeralo et al. 2007) as well as the population structure within *D. purpureum* (Mehdiabadi et al. 2009) and *D. giganteum* (Flowers et al. 2010). This sequence has enough resolution to distinguish between sister species in the dictyostelids (Schaap et al. 2006; Romeralo et al. 2007), and prior work with this sequence gives us information on the level of divergence among species accepted as different. We used the sequences to construct a gene tree of all wild clones. The sequences of the primers we used are listed in Table 2.2. These primers were previously used for phylogenetics in *D. purpureum* and *D. giganteum* (Mehdiabadi et al. 2009; Mehdiabadi et al. 2010).

We harvested DNA by collecting 5-10 fruiting bodies into 150 µl of a

| Name | Sequence 5'-3' |
|---|---|
| Sandie_A | AACCTGGTTGATCCTGCCAGT |
| 17S_r1 | AGATAATACAAGCTGAACTA |
| 17S_f2 | GCTCGTAGTTGAAGTTTAAG |
| 1340_r | TCGAGGTCTCGTCCGTTATC |
| 17S_f3 | CTAAGATATAGTAAGGATTG |
| 17S_r3 | ATGATCCATCCGCAGGTTCA |
| ITS_5.8_f1 | ACGGTAAAGTTAACG GATCG |
| ITS_5.8_r1 | ACTCTCACCCAAGTATAACT |
| ITS_5.8_f2 | AAACTGCGATAATTCACTTG |
| ITS_5.8_r2 | CCGTCTTCACTCGCCGTTAC |

**Table 2.2** Primers used in sequencing the 17S and 5.8S region of ribosomal DNA in *P. violaceum*

5% Chelex solution (Bio-Rad, Hercules, CA, USA), then added proteinase K to a concentration of 1.25 mg/ml and incubated this solution at 56°C for four hours then 98°C for 30 minutes.

We amplified this region with a polymerase chain reaction using Invitrogen's Platinum taq polymerase and 0.5 µM of each primer, using chelexed DNA as template. PCR cycling conditions were as follows: initial denaturation at 94°C for 5 min, followed by 30 cycles of 1 min denaturation at 94°C, 1 min annealing at 50°C, 1 min elongation at 72°C, followed by a final elongation at 72°C for 10 min. We sequenced all PCR products in both directions using Big Dye Terminators (Applied Biosciences, Foster City, CA, USA) and analyzed with an ABI Prism automated sequencer (Applied Biosciences, Foster City, CA, USA). We edited chromatograms and aligned contigs using the programs SeqMan (DNASTAR, Madison, WI, USA) and BioEdit (Hall, http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Sequences have been deposited in Genbank [HQ732139-HQ732228].

We followed procedures previously used in our group for phylogenetic analyses (Mehdiabadi et al. 2009). We included as outgroups *D. purpureum* and *D. citrinum* which are two group 4 dictyostelids (Genbank: *D. purpureum* DQ340386.1, *D. citrinum* DQ340385.1). We aligned sequences using ClustalW (Chenna et al.

2003). We developed a gene tree using Bayesian inference (Mr. Bayes, (Ronquist and Huelsenbeck 2003)). To determine the optimal nucleotide substitution model, we used Akaike information criteria (AIC) (Akaike 1974) and Bayesian information criteria (BIC) (Schwarz 1978), as implemented in ModelGenerator (Keane et al. 2006). A Generalized Time Reversible Model with a gamma distribution of mutations (GTR + Γ) was found to be the best model according to both AIC and BIC (data not shown). We used Mr. Bayes (Ronquist and Huelsenbeck 2003) to construct a gene tree and to estimate posterior probabilities for each node with parameters estimated based on the model recommended by ModelGenerator (Keane et al. 2006), the GTR + Γ model. The program ran four Metropolis-Coupled Markov chains for 1,600,000 generations following a burn-in period of 400,000 generations with sampling every 100 generations and beginning with a random tree. We looked at the average standard deviation of split frequencies to check convergence. By the 10,000 sampled tree, the average standard deviation of split frequencies had stabilized at ~0.011, and did not decrease in the following 6,000 sampled trees. We also checked convergence with Are We There Yet? (AWTY, http://ceb.scs.fsu.edu/awty) (Nylander et al. 2008). We used the 'compare' option to compare the posterior probabilities of clades from independent runs checking to make sure that the posterior probabilities of

the splits are the same for both independent runs. Nodes with posterior probabilities of less than 0.80 were collapsed.

We also generated a Maximum Likelihood tree using Garli (Zwickl 2006). We generated 500 bootstrap replicates. We used consense to generate a consensus tree with bootstrap support. Seqboot, dnaml, and consense are all part of the Phylip package (Felsenstein 2005).

C. Macrocyst formation experiments

When dictyostelid cells aggregate in response to starvation, there are two developmental pathways that the cells can take – the social, fruiting body stage or the sexual macrocyst stage as shown in Figure 2.1. Macrocyst formation is favored when cells are cultured in the dark, under liquid, and without phosphate (Nickerson and Raper 1973a). During macrocyst development, amoebae are attracted to the cAMP produced by the diploid fusion of two cells of different mating types. The attracted amoebae wall themselves in and are gradually consumed by the sexual cell which forms a giant cell that divides many times before germination when they release hundreds of recombined amoebae (Nickerson and Raper 1973a). Clones of the opposite sex and same species form macrocysts under appropriate conditions, though it is extremely difficult to get these macrocysts to hatch in the laboratory. This means that

macrocyst formation is only a partial test for true sexual compatibility.

To test for macrocyst formation, we incubated each clone both by itself and with each other clone tested under conditions favorable for macrocyst formation. We tested clones for macrocyst formation by plating spores on phosphate-free lactose peptone agar (1 g lactose, 1 g bactopeptone, 15 g agar, 1 l di$H_2O$) with *K. aerogenes* as a food source. We then flooded these plates with Bonner's standard salt solution (5.4 mM $CaCl_2$, 10 mM KCl, 5.1 mM NaCl), wrapped them in aluminum foil, and incubated in the dark for 3-5 days. After 5-7 days, we scored macrocysts as either present or absent for each treatment. When checked at later times (2-3 checks within 3-5 weeks) no additional macrocysts had formed. In mixes where no macrocysts had formed, cells usually reached aggregation stage and stopped or the cells simply died. In a few cases, cells made fruiting bodies or spores. For most mixes, the clones were divided into sets and all combinations of clones were mixed within that set. The sets were then replicated twice. If both clones were in two different sets, then that particular mix was performed 4 times (for example, QSvi9 and QSvi29).

D. Testing kin discrimination

To test for kin discrimination, we performed 13 reciprocal pairwise

mixes. For each mix, we fluorescently labeled two clones, and mixed each clone with unlabeled cells of the same clone and unlabeled cells of the other clone. We performed both the reciprocal mixes, to control for any effects of labeling, and mixes within the same clone to ensure that the cells were healthy. We allowed these four mixes to starve, aggregate and form fruiting bodies. We followed the same protocol as (Mehdiabadi et al. 2006).

Cells of each clone were grown up to log phase, split into two groups, one of which was labeled with 5-chloromethylfluorescein diacetate (CellTracker TM, Invitrogen, Carlsbad, CA, USA). These cells were then mixed together in the following fashion: labeled cells of the first clone mixed with unlabeled cells of the first clone, labeled cells of the second clone mixed with unlabeled cells of the second clone, labeled cells of the first clone mixed with unlabeled cells of the second clone, and labeled cells of the second clone mixed with unlabeled cells of the first. Additionally, we plated out the labeled and unlabeled cells of each clone alone as controls. We collected individual fruiting bodies from each treatment and counted the number of fluorescently labeled spores and the number of unlabeled spores to determine the proportion of each clone present in the fruiting body.

E. Statistics

To evaluate the extent of sorting in each fruiting body we calculated the average relatedness of the spores in each fruiting body assuming that each clone was completely related to itself (r=1) and completely unrelated to the other clone (r=0). Relatedness of the overall fruiting body is calculated as the proportion of labeled cells or spores squared plus the proportion of unlabeled cells or spores squared ($r=p^2+q^2$); that is, p of the cells are related by p to the other cells, and q of the cells are related by q. We calculated relatedness individually for each fruiting body and then averaged over all fruiting bodies. We measured sorting as a significantly higher relatedness in the experimental fruiting bodies than in control fruiting bodies.

Because the data were not normally distributed, we used Resampling Stats for Excel (Resampling Stats Inc., Arlington, VA, USA) to create a test. We calculated the test statistic [F= Variance (experimental)/Variance (control)] as the ratio of the average variance of the two experimental treatments divided by the average variance of the two control treatments. We sampled without replacement the dataset of the proportion of fluorescent spores of each individual fruiting body across the four treatments (two experimental and two control) 5000 times to determine the probability that a variance ratio as high as this observed

ratio could be obtained by chance (Mehdiabadi et al. 2006).

To test for geographic population structure, we ran an Analysis of Molecular Variance (AMOVA) using Arlequin 3.11 (Excoffier, Laval, and Schneider 2005) and used resampling (1023 permutations) to obtain significance values.

**Results**:

Sequence analysis and phylogeny:

We sequenced approximately 2500 bp of the 17S to 5.8S ribosomal DNA for 90 clones of *P. violaceum*. Out of the 90 clones sequenced, we identified 67 unique haplotypes. We aligned these sequences and used the two species *D. citrinum* and *D. purpureum* as outgroups. The resultant Bayesian gene tree is shown in Figure 2.2 with the support values from both the Bayesian gene tree and the maximum likelihood gene tree shown on the tree in Figure 2.2.i.

We find that the *P. violaceum* is split into six major groups, labeled A, B, C, D, E, and F on the phylogeny (Figure 2.2). Groups C, D, E, and F made up one of the basal branches of the phylogeny while the group A and group B made up the other two branches. The phylogeny shows some evidence for geographic population structure. Group A is comprised of all of the clones from Germany. All of the clones in group D were from

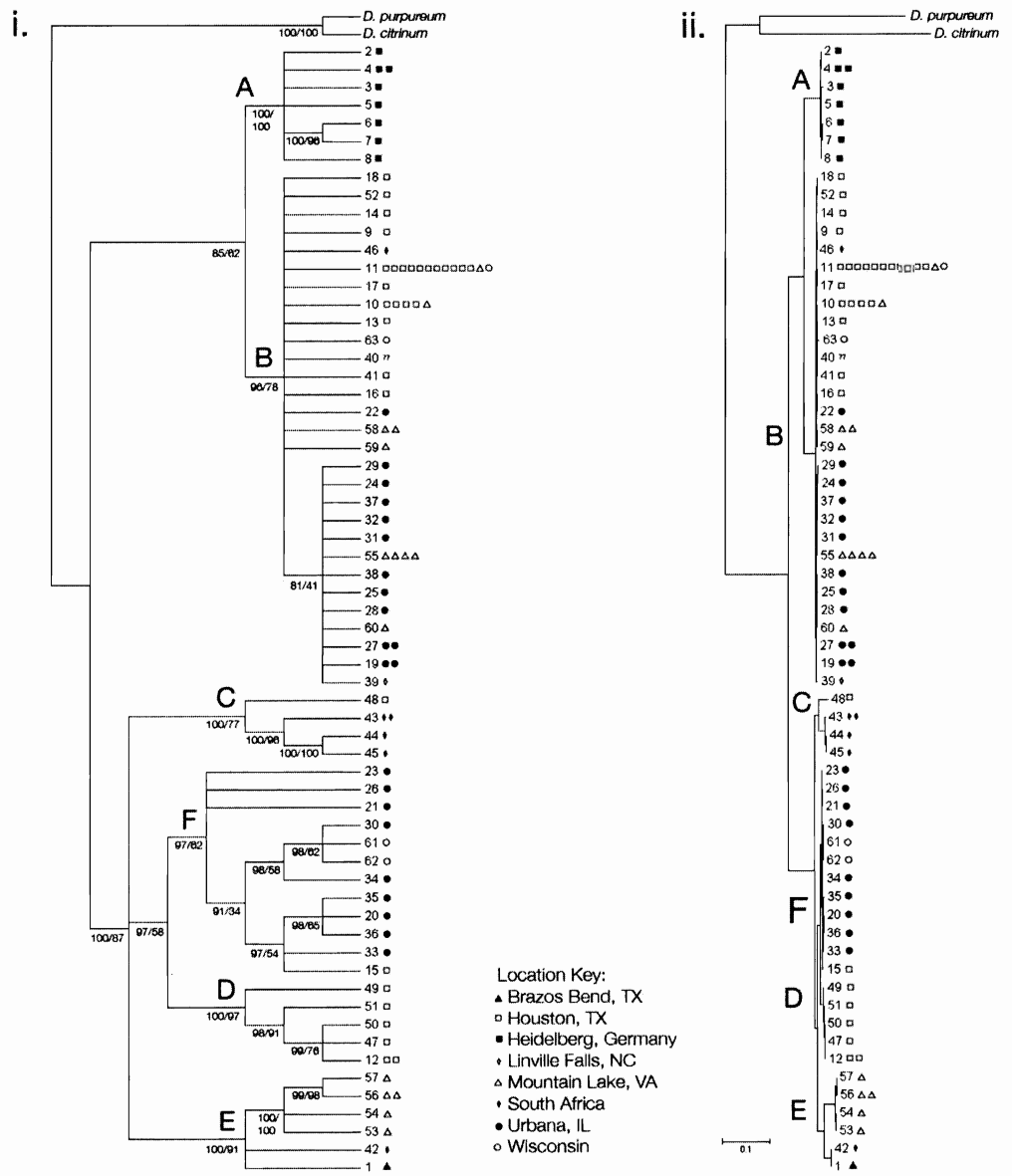



**Figure 2.2**. Bayesian gene tree based on ~2400 bp from 17S-5.8S RNA region of the ribosome of *P. violaceum* clones. *Dictyostelium purpureum* and *D. citrinum* were used as the outgroups. Each symbol represents one clone, and each branch represents one unique haplotype. The letters simply refer to different phylogenetic groups. i. Cladogram with nodes with Bayesian inference posterior probabilities of less than 0.95 collapsed. Numbers on the nodes are the Bayesian posterior probabilities and bootstrap values from the maximum likelihood analysis. ii. Phylogram.

Houston, TX, however Houston clones belonged to other groups as well. Most of the clones in groups C and E were from the same location (South Africa and Mountain Lake, VA respectively), however these locations also had clones that belonged to other groups. Not all phylogenetic divisions came from geographic structure. Clones in group B came from 6 of the 8 locations that we sampled.

Using a shorter sequence to enable the use of *D. laterosorum* as an outgroup results in a tree that is similar to the tree resulting from the full length sequence (Figure 2.3). Groups A and B still form a branch together. The node with groups C, D, E and F has been collapsed to a polytomy. All but one of the haplotypes in group C cluster as a group still. All but 2 haplotypes in group E cluster as a groups still. Group D remains unchanged. Group F has been collapsed entirely, and the groups C, D and E no longer show any relationship with one another.

To see if population structure was due to geographic distance, we calculated the Analysis of Molecular Variance (AMOVA) using the 17S to 5.8S sequence. Like Fst, the AMOVA is a measure of the population variance, however the AMOVA also incorporates the degree of difference (mutations) between alleles (Excoffier, Smouse, and Quattro 1992). The AMOVA showed that 7.75% of the genetic variation observed was between geographically delimited populations, and the variance was

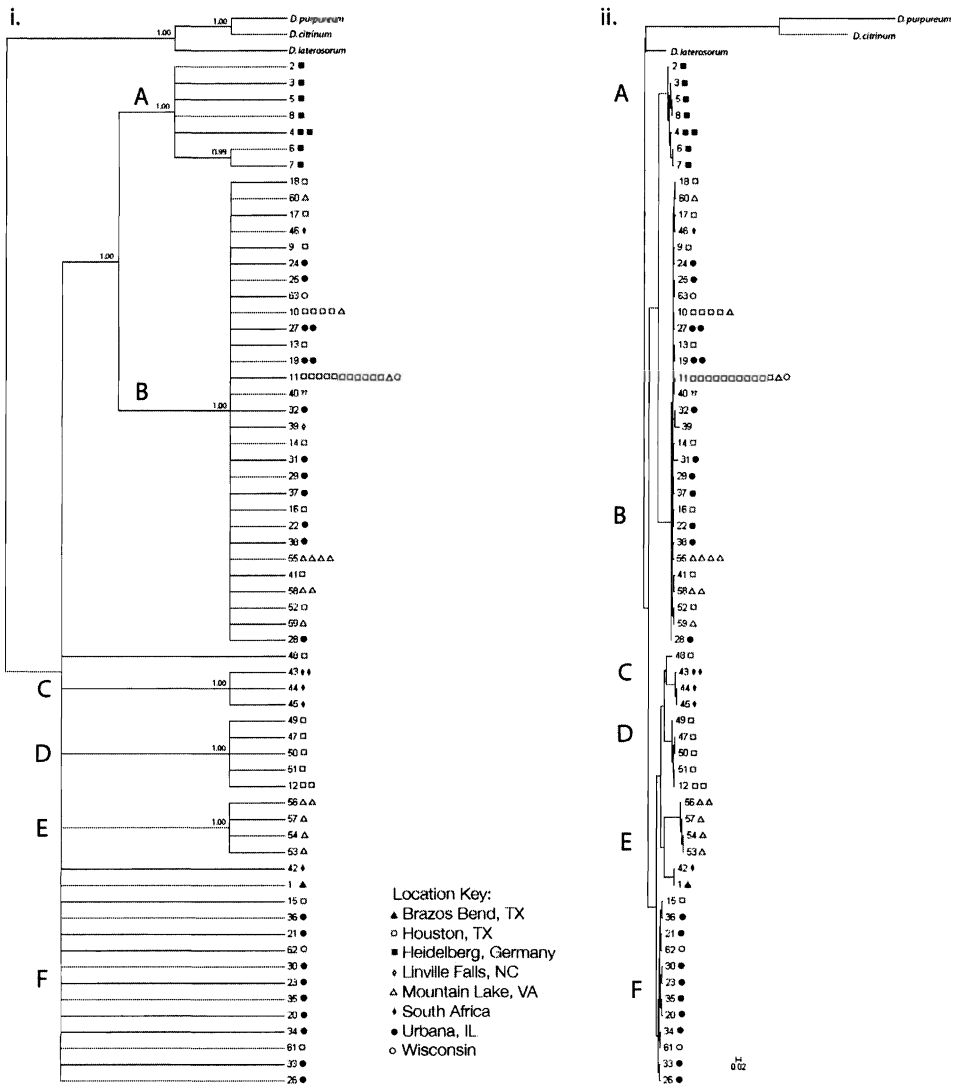



**Figure 2.3** Bayesian gene tree based on ~1700 bp from 17S ribosomal RNA region of *P. violaceum*. Genbank accessions *D. purpureum* DQ340386.1, *D. citrinum* DQ340385.1, and *D. laterosorum* AM168046.1 were used as outgroups. The tree was constructed as detailed in the methods, with the constraint that all of the outgroups had to group together. Each symbol represents one clone, and each branch represents one unique haplotype. The letters simply refer to different phylogenetic groups. i. Cladogram with nodes with Bayesian inference posterior probabilities of less than 0.95 collapsed. Numbers on the nodes are the Bayesian posterior probabilities. ii. Phylogram.

significant ($p<0.00001$).

To look at the level of divergence between the phylogenetic groups, we used the 17S sequence to calculate pairwise distances (base substitutions per site) between all of the clones. We used MEGA4 to calculate the distances using the Maximum Composite Likelihood method (Tamura, Nei, and Kumar 2004; Tamura et al. 2007). Using just the 17S sequence, we calculated the average pairwise distance between clones of *P. violaceum* and *D. laterosorum* (clone AE4). These distances ranged from 0.013 to 0.019 (data not shown). Between the two major groups (A/B and C/D/E/F), pairwise distances ranged from 0.010 to 0.021 (data not shown). Within these two major groups, distances ranged from 0.000 to 0.010. Within each of the six groups, the maximum pairwise distances ranged from 0.000 (groups B, C, and F) to 0.010 (group E). Within groups, all clones had a minimum pairwise distance of 0.000. The average pairwise distances between some clones are analogous to the average pairwise distances between clones of *P. violaceum* and *D. laterosorum,* which are clearly different species. This suggests that *P. violaceum* has species level diversity.

Macrocyst matings:

To further look at speciation in *P. violaceum*, we also examined

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0/21 | 0/24 | 0/1 | 0/7 | 0/7 | 0/3 |
| B | | 46/287 | 1/101 | 0/75 | 1/33 | 5/87 |
| C | | | 0/6 | 0/26 | 4/17 | 2/14 |
| D | | | | 6/12 | 1/22 | 2/13 |
| E | | | | | 3/6 | 0/7 |
| F | | | | | | 16/66 |

**Table 2.3**. Macrocyst production in *P. violaceum.* The letters refer to the phylogenetic groups from Figure 1.2. A total of 835 unique mixes were attempted between two clones. In each category, matings are listed as successful mixes / attempted mixes. Each mix between any two individuals was performed at least twice. When these gave conflicting results, another replicate mix was performed and the majority result was counted.

macrocyst formation (Table 2.3). No individuals from group A had any successful matings, either with other members of group A or members of other groups. In group B, we observed two different mating types: we had two clones of one mating type and 23 clones of the other mating type. With one exception, all matings between clones of compatible mating types resulted in macrocyst formation. In the other groups C, D, E, and F, the mating types were not as clear because of triads of clones where all three clones would mate with each other in pairwise combinations, leading to uncertainty about the number of sexes and whether clones might be bisexual. Groups D and E both had two clearly defined mating types, and each clone made the sexual form called a macrocyst when paired with a member of the same group but opposite mating type. One clone from group E also mated with group C, though group C did not mate at all within itself, perhaps because our sample did not include compatible sexes. Clones from group F mated between themselves as well as with clones from other groups. We performed multiple replicates of each set of mixes. While most mixes were consistent between replicates, some mixes (16 out of 835 mixes) formed macrocysts only in some of the replicates. In these cases, the result (macrocyst formation or no macrocyst formation) that was found in the majority of replicates was used, to rule out the possibility of contamination or occasional selfing.

Overall, we saw complete mating within group B (all clones of mating type A mated with all of mating type B), and a high degree of mating within the other major clade of groups C, D, E, and F, with a few matings between two major clades. Between the groups C, D, E, and F there was some mating between the different groups though matings were not consistent enough to be able to assign mating types to clones and thus diagnose the thoroughness of mating success.

We are hesitant to categorize the rate of successful matings between because of the variation in mating types throughout the dictyostelids. *D. discoideum* has at least two different mating types, with a bisexual mating type in addition (Urushihara H 1992). In *D. rosarium*, at least three mating types are present (Chang MT and KB 1981). Group B of *P. violaceum* had two well-defined mating types, and all clones mated when paired with a member of the opposite mating type of Group B. However, we were unable to identify the number of mating types within the other groups of *P. violaceum*. Because of this, we are not confident in estimating the number of matings that could occur based on mating types and this is necessary to compare the rates of observed matings to the possible matings.

Kin discrimination:

We looked at the influence of phylogenetic relationships between individuals on cooperation. To do this, we related genetic distance to the degree of sorting between clones. Both distantly related clones and clones from different species should be less likely to cooperate in forming fruiting bodies. All but three pairwise mixes showed significant sorting in comparison to controls (Figure 2.4, Resampling stats). These three mixes were all between members of group B (Figure 2.4). While the rest of the mixes were all statistically significant from the controls, some of the mixes between groups C-F showed incomplete sorting (relatedness <1). The average relatedness of the pairwise mixes was 0.8 (0.05 std error).

**Discussion:**

Our data suggest that *P. violaceum* might contain cryptic species. There are 6 groups (A, B, C, D, E, and F) with the most basal split being between the groups A/B and C/D/E/F. Though the genetic distances between these two main groups are relatively small (0.010-0.021), they are larger than the differences between accepted species in the dictyostelids using the same part of the 17S sequence, for example between *D. citrinum* (OH494) and *D. dimigraformum* (AR5b), 0.009; *D. clavatum* (TNS-C-189) and *D. longosporum* (TNS-C-109), 0.003; *D. mucoroides* (TNS-C-114) and *D. sphaerocephalum* (GR11), 0.001 or

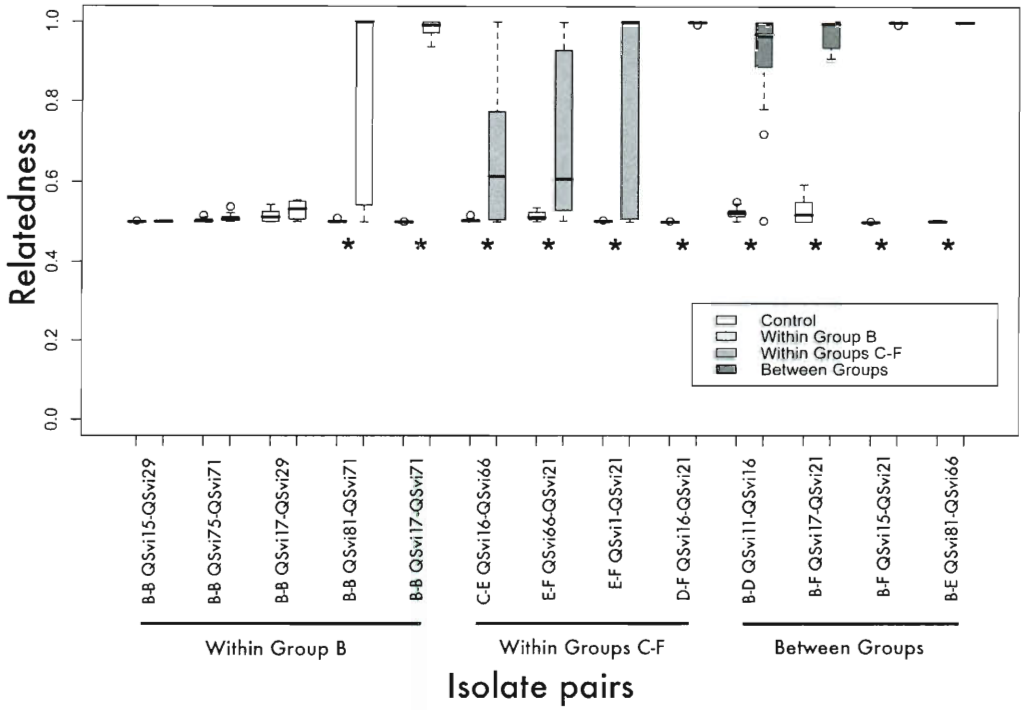



**Figure 2.4.** Box and whisker plot of relatedness of individual clones in fruiting bodies. Relatedness is calculated as the proportion of labeled cells squared plus the proportion of unlabeled cells squared ($r=p^2+q^2$), which assumes r=1 to clonemates and r=0 to non-clonemates. Relatedness was calculated for each fruiting body individually. Relatedness varies from 0.5 (complete mixing) to 1 (sorting). Groups refer to the six lettered groups on the phylogenetic tree. Resampling stats were used to compare the relatedness of the control mix (labeled and unlabeled cells of the same clone) to the experimental (labeled and unlabeled cells from different clones). Error bars are standard error of the mean. * the control mix relatedness is lower than the experimental $p< 0.001$. See methods for details.

between *D. brunneum* (WS700) and *D. giganteum* (WS589) 0.004 (Mehdiabadi et al. 2009)). The 17S to 5.8S sequence is relatively conserved compared to other genes in a wide variety of organisms (for example (Bailey and Andersen 1999; Evans, Wortley, and Mann 2007; Daniels et al. 2009)). This sequence has enough resolution to distinguish between sister taxa in the dictyostelids (Schaap et al. 2006; Romeralo et al. 2007).

Furthermore, the variation between geographical populations accounts for less than 10% of the total variation. Many haplotypes are found in more than one geographic population. In addition, group B is cosmopolitan, with individuals from almost every geographic location belonging to this group. This suggests that the population structure in *P. violaceum* is not due to geographic constraints alone. The major differences that we see in the 17S seem for the most part consistent with species-level differences, with the species often occurring in the same areas.

By and large, the division between the two groups was reinforced by our mating experiments. Like Clark (Clark, Francis, and Eisenberg 1973; Clark 1974), we found two groups of non-interbreeding individuals; however, we observed a few instances of mating between the groups. This leaves open the possibility for some gene exchange between

different groups should those macrocysts actually be able to germinate. Unlike the model organism, *D. discoideum*, *P. violaceum* reported germination rates have been upwards of 50% (Nickerson and Raper 1973b). Examining the germination of our between-group macrocysts, which would require prolonged ageing of macrocysts (Nickerson and Raper 1973b), would be a fruitful line of research for the future.

The split between the two groups is also apparent when looking at cooperation during fruiting body formation. We have found that only clones from group B exhibit strong mixing and cooperation with other group B clones in forming the fruiting. When the clones were from different phylogenetic groups sorting was more complete. Both the phylogenetic diversity and the behavioral changes suggest that there may be at least two different morphologically identical sister species in *P. violaceum*. Both lines of evidence are consistent with the same division and the variation that we observed in phylogenetic structure affects the behavior that we observe in the social stage.

Relatedness allows altruism to be beneficial if the altruistic acts are directed towards relatives. Because clones from two different species are not related, there should be selection for species discrimination. Because we are unsure of the exact nature of the relationship between individual clones, we use the term kin discrimination rather than species

discrimination. In *D. discoideum,* the further the genetic distance between clones, the greater the propensity for kin discrimination to occur (Ostrowski et al. 2008). In our study, a few clones cooperated with each other to form chimeric fruiting bodies, but most clones tested sorted out to form mostly clonal fruiting bodies. All the clones that cooperate with each other were in the same group (B). This fits with the idea of kin selection, with only closely related clones cooperating, though we do not have information on exact values for the other half of kin selection: the relative costs and benefits of cooperation. Benefits of larger groups are likely to include lower proportions of cells destined for stalk relative to spore, and ability to move greater distances, while costs center on becoming a sterile stalk cell.

Previous studies on kin discrimination in the dictyostelids have given differing results depending on the species used. Kin discrimination has also been investigated in *Dictyostelium discoideum*, *D. purpureum* and *D. giganteum*. Clones of *D. discoideum* exhibit kin discrimination with more distantly related clones sorting more than clones that are more closely related (Ostrowski et al. 2008). *D. purpureum* shows kin discrimination as well (Mehdiabadi et al. 2006). In *D. giganteum*, some genetically distinct clones exhibit kin discrimination while others do not (Kaushik, Katoch, and Nanjundiah 2006). The question of whether *D.*

*giganteum* is one species worldwide with varying levels of kin discrimination or multiple cryptic species has not been resolved, but North American clones show no differentiation (Mehdiabadi et al. 2010). Our results show that *P. violaceum* exhibits kin discrimination; like the other dictyostelids, clones from different cryptic groups within *P. violaceum* sort to form clonal fruiting bodies while closely related clones sometimes cooperate to form chimeric fruiting bodies.

Most of the dictyostelids have been identified and distinguished on the basis of morphology. An exception is recent work on *Polysphondylium pallidum* and its sister species *P. album* (Kawakami and Hagiwara 2008) as well as *D. ibericum* (Romeralo, Baldauf, and Cavender 2009). Romeralo, Baldauf, and Cavender (Romeralo, Baldauf, and Cavender 2009) used morphology to identify a new species, and molecular phylogenetics to place that species within the dictyostelids. Kawakami and Hagiwara (Kawakami and Hagiwara 2008) use a combination of mating type and morphological characters to redefine these two species. They show that there are three groups, one that matches the *P. pallidum* type specimen and mates with *P. pallidum* strains, one that matches the *P. album* type specimen and mates with *P. album* strains, and one that matches neither exactly and mates with neither. The relationship of the third group to *P. pallidum* and *P. album* remains unclear. These recent

studies make it clear that relying on morphology alone to dictate species boundaries is not sufficient, and mating type analysis and molecular work is needed to correctly identify species boundaries and the relationships between species.

Improperly identifying cryptic species also affects biodiversity metrics as well as estimates of geographical distributions. By identifying all members of a cryptic species complex as the same species, biodiversity is underestimated and geographic distributions are overestimated. In the identification of protists this can be especially difficult because of a lack of distinguishing morphological characteristics (Slapeta, Lopez-Garcia, and Moreira 2006). The difficulty of correctly identifying cryptic species has contributed to debate on protist biogeography. Finlay (Finlay 2002) suggests that there is something fundamentally different about microorganisms, including protists, such as higher rates of migration and lower rates of speciation that causes them to be more cosmopolitan than larger organisms. Foissner (Foissner 2006) suggests that more endemic species are present in part because of molecularly distinct but morphologically similar species that are endemic but are classified as a single cosmopolitan species.

Molecular sequence data has identified many cryptic protist species. Most of the cases involve apparently cosmopolitan species that

are actually comprised of geographically restricted cryptic species (Foissner 2006). However, this is not always the case. *Aspergillus fumigatus* is composed of several cryptic species, but rather than being geographically isolated species, at least one species is globally distributed (Pringle et al. 2005). Similarly, phylogenetic analyses divide the desert truffle *Terfezia boudieri* into 3 morphologically identical groups (Ferdman et al. 2009). All collections were made in a roughly 50 $km^2$ region of the Negev desert. While these 3 species are endemic, they do have overlapping ranges, and there is the possibility that all 3 might have the same range. In addition, the ectomycorrhizal fungus, *Tricholoma scalpturatum* shows two distinct groups (Carriconde et al. 2008). Roughly half of the genetic variance was found within populations and half of it was found to be between populations. At short ranges, populations were sometimes structured, but both groups were represented over the entire sampling range (France to Sweden). These previous studies suggest that cryptic species are not always endemic subgroups of a cosmopolitan morphotype. If *P. violaceum* is in fact composed of two morphologically indistinguishable species that are globally distributed, they support Foissner's (Foissner 2006) idea that while globally distributed cosmopolitan species exist, morphologically identical endemic species are also present.

**References:**

Akaike, H. 1974. New look at statisitical-model identification. Ieee Transactions on Automatic Control **AC19**:716-723.

Bailey, J. C., and R. A. Andersen. 1999. Analysis of clonal cultures of the brown tide algae *Aureococcus* and *Aureoumbra* (Pelagophyceae) using 18S rRNA, *rbc*L, and rubisco spacer sequences. Journal of Phycology **35**:570-574.

Briscoe, D. A., A. A. Gooley, R. L. Bernstein, G. M. McKay, and K. L. Williams. 1987. Genetic diversity in cellular slime-molds - allozyme electrophoresis and a monoclonal-antibody reveal cryptic species among *Dictyostelium discoideum* strains. Genetics **117**:213-220.

Carriconde, F., M. Gardes, P. Jargeat, J. Heilmann-Clausen, B. Mouhamadou, and H. Gryta. 2008. Population evidence of cryptic species and geographical structure in the cosmopolitan ectomycorrhizal fungus, *Tricholoma scalpturatum*. Microbial Ecology **56**:513-524.

Cavender, J. C., J. Cavender-Bares, and H. R. Hohl. 1995. Ecological distribution of cellular slime molds in forest soils of Germany. Botanica Helvetica **105**:199-219.

Chang MT, and R. KB. 1981. Mating types and macrocyst formation in *Dictyostelium rosarium*. Journal of Bacteriology **147**:1049-1053.

Chenna, R., H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Research **31**:3497-3500.

Clark, M. A. 1974. Syngenic divisions of cellular slime-mold *Polysphondylium violaceum*. Journal of Protozoology **21**:755-757.

Clark, M. A., D. Francis, and Eisenberg. 1973. Mating types in cellular slime molds. Biochemical and Biophysical Research Communications **52**:672-678.

Daniels, S. R., M. D. Picker, R. M. Cowlin, and M. L. Hamer. 2009. Unravelling evolutionary lineages among South African velvet worms (Onychophora: *Peripatopsis*) provides evidence for widespread cryptic speciation. Biological Journal of the Linnean Society **97**:200-216.

Evans, K. M., A. H. Wortley, and D. G. Mann. 2007. An assessment of potential diatom "barcode" genes (*cox*1, *rbc*L, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). Protist **158**:349-364.

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online **1**:47-50.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes - Application to human mitochondrial-DNA restriction data. Genetics **131**:479-491.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Ferdman, Y., Y. Sitrit, Y. F. Li, N. Roth-Bejerano, and V. Kagan-Zur. 2009. Cryptic species in the *Terfezia boudieri* complex. Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology **95**:351-362.

Feulner, P. G. D., F. Kirschbaum, C. Schugardt, V. Ketmaier, and R. Tiedemann. 2006. Electrophysiological and molecular genetic evidence for sympatrically occuring cryptic species in African weakly electric fishes (Teleostei : Mormyridae : *Campylomormyrus*). Molecular Phylogenetics and Evolution **39**:198-208.

Fey, P., P. Gaudet, K. E. Pilcher, J. Franke, and R. L. Chisholm. 2006. dictyBase and the dicty stock center. Methods in Molecular Biology:51-74.

Finlay, B. J. 2002. Global dispersal of free-living microbial eukaryote species. Science **296**:1061-1063.

Flowers, J. M., S. I. Li, A. Stathos, G. Saxer, E. A. Ostrowski, D. C. Queller, J. E. Strassmann, and M. D. Purugganan. Variation, sex, and social cooperation: molecular population genetics of the social amoeba Dictyostelium discoideum. PLoS Genet **6**:e1001013.

Foissner, W. 2006. Biogeography and dispersal of micro-organisms: A review emphasizing protists. Acta Protozoologica **45**:111-136.

Hagiwara, H. 1990. Altitudinal distribution of dictyostelid cellular slime molds in the Langtang Valley Nepal of Central Himalayas. Reports of the Tottori Mycological Institute:191-198.

Hagiwara, H. 2000. Dictyostelids in the region around the Seto Inland Sea, Japan. Memoirs of the National Science Museum (Tokyo):77-81.

Kaushik, S., B. Katoch, and V. Nanjundiah. 2006. Social behaviour in genetically heterogeneous groups of *Dictyostelium giganteum*. Behavioral Ecology and Sociobiology **59**:521-530.

Kawakami, S. I., and H. Hagiwara. 2008. A taxonomic revision of two dictyostelid species, *Polysphondylium pallidum* and *P. album*. Mycologia **100**:111-121.

Keane, T. M., C. J. Creevey, M. M. Pentony, T. J. Naughton, and J. O. McInerney. 2006. Assessment of methods for amino acid matrix

selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. Bmc Evolutionary Biology **6**.

Landolt, J. C., J. C. Cavender, S. L. Stephenson, and E. M. Vadell. 2008. New species of dictyostelid cellular slime moulds from Australia. Australian Systematic Botany **21**:50-66.

Maeder, A., A. Freitag, and D. Cherix. 2005. Species- and nestmate brood discrimination in the sibling wood ant species *Formica paralugubris* and *Formica lugubris*. Annales Zoologici Fennici **42**:201-212.

Mehdiabadi, N. J., C. N. Jack, T. T. Farnham, T. G. Platt, S. E. Kalla, G. Shaulsky, D. C. Queller, and J. E. Strassmann. 2006. Kin preference in a social microbe - Given the right circumstances, even an amoeba chooses to be altruistic towards its relatives. Nature **442**:881-882.

Mehdiabadi, N. J., M. R. Kronforst, D. C. Queller, and J. E. Strassmann. 2009. Phylogeny, reporductive isolation and kin recognition in the social amoeba *Dictyostelium purpureum*. Evolution **63**:542-548.

Mehdiabadi, N. J., M. R. Kronforst, D. C. Queller, and J. E. Strassmann. 2010. Phylogeography and sexual macrocyst formation in the social amoeba *Dictyostelium giganteum*. BMC Evolutionary Biology **10**.

Nickerson, A. W., and K. B. Raper. 1973a. Macrocysts in the life cycle of the Dictyosteliaceae. Part 1 Formation of the macrocysts. American Journal of Botany **60**:190-197.

Nickerson, A. W., and K. B. Raper. 1973b. Macrocysts in life-cycle of Dictyosteliaceae. 2. Germination of macrocysts. American Journal of Botany **60**:247-254.

Nylander, J. A. A., J. C. Wilgenbusch, D. L. Warren, and D. L. Swofford. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics **24**:581-583.

Ostrowski, E. A., M. Katoh, G. Shaulsky, D. C. Queller, and J. E. Strassmann. 2008. Kin Discrimination Increases with Genetic Distance in a Social Amoeba. Plos Biology **6**:2376-2382.

Pringle, A., D. M. Baker, J. L. Platt, J. P. Wares, J. P. Latge, and J. W. Taylor. 2005. Cryptic speciation in the cosmopolitan and clonal human pathogenic fungus *Aspergillus fumigatus*. Evolution **59**:1886-1899.

Raper, K. B. 1984. The dictyostelids. The dictyostelids.:i-xi, 1-453.

Romeralo, M., S. L. Baldauf, and J. C. Cavender. 2009. A new species of cellular slime mold from southern Portugal based on morphology, ITS and SSU sequences. Mycologia **101**:269-274.

Romeralo, M., R. Escalante, L. Sastre, and C. Lado. 2007. Molecular systematics of dictyostelids: 5.8S Ribosomal DNA and internal transcribed spacer region analyses. Eukaryotic Cell **6**:110-116.

Romeralo, M., and C. Lado. 2006. Dictyostelids from Mediterranean forests of the south of Europe. Mycological Progress **5**:231-241.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

Saez, A. G., and E. Lozano. 2005. Body doubles. Nature **433**:111-111.

Schaap, P., T. Winckler, M. Nelson, E. Alvarez-Curto, B. Elgie, H. Hagiwara, J. Cavender, A. Milano-Curto, D. E. Rozen, T. Dingermann, R. Mutzel, and S. L. Baldauf. 2006. Molecular phylogeny and evolution of morphology in the social amoebas. Science **314**:661-663.

Schwarz, G. 1978. ESTIMATING DIMENSION OF A MODEL. Annals of Statistics **6**:461-464.

Slapeta, J., P. Lopez-Garcia, and D. Moreira. 2006. Global dispersal and ancient cryptic species in the smallest marine eukaryotes. Molecular Biology and Evolution **23**:23-29.

Sutherland, J. B., and K. B. Raper. 1978. Distribution of cellular slime molds in Wisconsin prairie soils. Mycologia **70**:1173-1180.

Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution **24**:1596-1599.

Tamura, K., M. Nei, and S. Kumar. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences of the United States of America **101**:11030-11035.

Urushihara H. 1992. Sexual development of cellular slime molds. Development Growth and Differentiation **34**:1-17.

Vadell, E. M. 2000. Dictyostelids (Eumycetozoa) from soils of Punta Lara, Province of Buenos Aires, Argentina. Revista Argentina de Microbiologia **32**:89-96.

Vadell, E. M., and J. C. Cavender. 1998. *Polysphondylium* from forest soils of Tikal, Guatemala. Mycologia **90**:715-725.

Wulff, J. L. 2006. Sponge systematics by Starfish: Predators distinguish cryptic sympatric species of Caribbean fire sponges, *Tedania ignis* and *Tedania klausi* n. sp (Demospongiae, Poecilosclerida). Biological Bulletin **211**:83-94.

Yeh, Z. Y. 2003. Biodiversity inventory of dictyostelid cellular slime molds in Taiwan. Mycotaxon **86**:103-110.

Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin.

# CHAPTER 3

## Linkage disequilibrium in microsatellites: a case study in *Dictyostelium discoideum*

**Abstract:**

Accurately measuring linkage disequilibrium is important in identifying recombination as well as for doing association mapping. D' is a commonly used measure of linkage disequilibrium. However, previous studies have shown that it can be biased. We used the model organism *Dictyostelium discoideum* to investigate how the measure of linkage disequilibrium, D', is influenced by sample size and number of alleles. We used a dataset of 100 microsatellite loci from 24 different isolates of *D. discoideum*. Loci that were close together have higher D' values than those on different chromosomes. However average D' values were remarkably high, even in comparisons between chromosomes where simple meiosis should reduce disequilibrium. We investigated several different approaches to calculating linkage disequilibrium, using computer simulations to assess the efficacy of different measures under a variety of conditions. While D' should range from 0 to 1 in theory, the simulations show that for the parameters in our populations, the minimum D' is much higher. D' was heavily biased, both in our sample and in simulations, by sample size, the number of possible haplotypes, and allele frequency. We used two different corrections to D' to see if this bias could be alleviated. A correction based on resampling ($D'_r$) reduced bias due to the number of possible haplotypes. A correction based on D' values of non-syntenic loci

($D'_c$) also reduced bias due to the number of haplotypes as well as reducing bias due to sample size. These results show that uncorrected D' values may not be good indicators of linkage disequilibrium, particularly when dealing with microsatellite loci. Corrections may only partially alleviate the bias due to sample size and the number of possible haplotypes, but corrections are necessary in situations where linkage disequilibrium measurements are being compared, such as the identification of the best marker of a disease state.

**Introduction:**

Linkage disequilibrium is the non-random association of alleles at two different loci (Lewontin 1988). Linkage disequilibrium has long been an important tool in population genetics and molecular biology (Mueller 2004). One of linkage disequilibrium's common uses is in association mapping. Association mapping is a method for the identification of genetic markers that predict phenotypic traits (Slatkin 2008). Association mapping relies on linkage disequilibrium to uncover these associations (Slatkin 2008). In population genetics, a lack of linkage disequilibrium can be used to infer the presence of recombination and sex (Schurko et al. 2009).

Accurately gauging linkage disequilibrium is therefore key to understanding the genetic architecture underlying phenotypic traits,

population structure because these areas require comparisons between estimates of linkage disequilibrium (Kruglyak 1999). A commonly used measure of linkage disequilibrium is D' (Lewontin 1964a). For two loci, D' is calculated by subtracting the expected frequency of each haplotype from the observed frequency of the haplotype and then normalizing by the maximum possible D value. The proportion of the sample that has the haplotype ij is denoted as $x_{ij}$, while $p_i$ and $q_j$ are the proportions of those alleles in the population.

$$D_{ij} = x_{ij} - p_i q_j$$

$$D'_{ij} = \frac{D_{ij}}{D_{\max}}$$

When $D_{ij} < 0$, $D_{\max} = \min\left[p_i q_j, (1-p_i)(1-q_j)\right]$

When $D_{ij} > 0$, $D_{\max} = \min\left[p_i(1-q_j), (1-p_i)q_j\right]$

$$D' = \sum_i \sum_j p_i q_j \left|D'_{ij}\right|$$

Dividing by $D_{max}$ standardizes D' so that it runs from 0 to 1 with 0 being equilibrium and 1 being complete disequilibrium. However, previous reports show that multiallelic data, such as microsatellites, tend to have higher D' values than biallelic data, such as single nucleotide polymorphisms (SNPs) (Abecasis et al. 2001; Slate and Pemberton 2007), even when the same individuals and genomic regions are being used

(Varilo et al. 2003). D' significantly increases as the heterozygosity of the microsatellite loci increases (McRae et al. 2002).

While these patterns could be due to differences in mutation rates, they could also be due to the way that linkage disequilibrium is calculated. When two loci are in linkage equilibrium, all possible combinations of alleles are present in the population at their expected frequencies. When both loci are biallelic, as is the case with SNPs, there are only 4 possible haplotypes. As the number of alleles at each locus is increased, the number of haplotypes that must been seen in the population in order for the loci to be considered in equilibrium increases. This increase in the number of haplotypes means that each haplotype is proportionately less likely to appear in the sample used to calculate D'. More generally, the lower the frequency of a haplotype, the greater the effect of chance in causing deviations of the observed frequency from the expected, and positive and negative deviation do not cancel, because the absolute value is taken.

Numerous ways to test the statistical significance of D' values have been proposed (Devlin et al. 2001; Zapata et al. 2001; Zaykin et al. 2008). However, statistical significance does not give an indication of the strength of the linkage disequilibrium. When comparing D' values, the comparison should be between the D' values themselves, or some related statistic, not

the results of any statistical test. It is desirable to find a standardization to eliminate the biases caused by sample size, allele number and allele frequency.

Many corrections to D' have been proposed (Weir and Cockerham 1978; Hedrick 1987; Devlin et al. 2001). Reducing multiallelic systems to biallelic ones does not work because the pooling can reduce the power to test for linkage disequilibrium (Weir and Cockerham 1978). Randomly resampling the dataset should break up any association of alleles in the dataset, rendering a dataset with no linkage disequilibrium but preserving sample size and allele number (Devlin et al. 2001). This should then reflect the minimum value of D' for that comparison in that dataset. Because any bias due to sample size or allele number affects the minimum value, this should help normalize D'. However, it is unclear if these corrected D' values are free from bias. Here we test this resampling correction as well as a novel correction factor.

In this paper we propose and test another correction for the bias, which we use to compare LD across a chromosome. It also works by subtracting another D' that includes bias, so that the bias should be eliminated or at least reduced in the difference. In this case, for the D' of any two loci, we subtract the average D' for these two loci to all loci on other chromosomes. It thus measures the excess D' for this pair above

and beyond the average value that each has with unlinked loci, each of which is subject to any bias due to sample size or variability at the focal locus.

Our aim is to evaluate the effects of the bias that sample size and the number of alleles have on D' and to test corrections to see if these corrections remove bias from D'. For this purpose we used a dataset of microsatellite genotypes of the social amoeba *Dictyostelium discoideum,* which first brought our attention to the biases caused by sample size and allele number on measurements of linkage disequilibrium. *D. discoideum* is one of the 13 model organisms that have been recognized by the NIH (http://www.nih.gov/science/models). It has been used to look at multicellular development (Devreotes 1989; Kessin 2001) and immunology (Williams et al. 2006) as well as social evolution (Strassmann et al. 2000). The role of sex has not been clear in the lifecycle of this haploid organism (Raper 1984). There is a sexual stage called a macrocyst, but under laboratory conditions, germination efficiency is remarkably low, with less than 1% of all macrocysts germinating (Nickerson and Raper 1973; Francis and Eisenberg 1993; Francis 1998). In some cases, these progeny show an excess of the parental genotypes (Francis and Eisenberg 1993) while others show recombinant genotypes (Francis 1998). Recent work using SNPs shows that in *D. discoideum* clones

collected from the wild, linkage disequilibrium decays rapidly with distance between loci, as expected if recombination occurs (Flowers et al. 2010).

Flowers et al. (2010) sequenced a set of 24 clones of *D. discoideum* for 184 SNPs. They used $r^2$ (Hill and Robertson 1968) and the 4 gamete test (Hudson and Kaplan 1985) to look for linkage disequilibrium and recombination on chromosome 4. They found that $r^2$ decays to the baseline level of $r^2$ values obtained from loci on separate chromosomes between 10 and 25 kb. With the 4 gamete test, they found a minimum of 14 recombination events on chromosome 4. This provides molecular evidence that *D. discoideum* does have sex in nature.

We use microsatellites genotyped in the same set of clones used by Flowers et al. (2010) to look at linkage disequilibrium and bias in D'. We also used computer simulations to show how sample size and the number of alleles per locus influence D'. We present two different corrections, and compare their efficiency at normalizing D'.
These corrections should enable the use of microsatellite markers in studies that compare levels of linkage disequilibrium, such as association mapping.

**Methods:**

Clones:

We used 24 clones of *D. discoideum*. Thirteen of these clones were collected from Mountain Lake, VA and 11 were collected from other parts of the US. One clone was collected from each of the following locations: Forrest City, AR, Carthage, TX, Houston Arboretum, TX, Bloomington, IN, Land Between the Lake, KY, Linden TX, Indian Gap, TN, Mt. Greylock, MA, Linville Falls, NC, Effingham, IL, and St. Louis, MO. These were cultured on SM medium (glucose 10 g, Oxoid Bactopeptone 10 g, yeast extract 1 g, $MgSO_4$ 1g, $KH_2PO_4$ 1.9 g, $K_2HPO_4$ 0.6 g, Bacto agar 20 g, in 1 L di$H_2O$) with *Klebsiella aerogenes* as a food source (Sussman 1966).

Microsatellites:

We selected 100 microsatellites for analysis. These 100 microsatellites were scattered throughout the genome. We sequenced these 100 microsatellites using primers that we generated ourselves. Primers were designed using Primer3 (Rozen and Skaletsky 2000). We added a 23 bp M13-tail (5'-AGGGTTTTCCCAGTCACGACGTT-3') to the forward primer to label sequences with a single fluorescent tag (Chbel et al. 2002).

We harvested DNA by collecting spores into a 5% Chelex solution (Bio-Rad, Hercules, CA, USA). We then added proteinase K to a concentration of 1.25 mg/ml. We incubated this solution at 56°C for four

hours then 98°C for 30 minutes. We amplified targeted microsatellites using the polymerase chain reaction (Saiki et al. 1985). PCR reagents for each 20 µl reaction included: 5.75 µl of DI $H_2O$, 2 µl of 10x buffer (100 mM Tris pH 8.3, 500 mM KCl, and 1% Triton X-100), 1.2 µl of 25 mM $MgCl_2$, 0.6 µl of 2.5 mM dNTP, 0.25 µl of 1 mg/ml BSA, 2 µl of *D. discoideum* DNA, 8 µl of 1.25 µM primer, and 0.2 µl of 5 U/µl BIOTAQ DNA polymerase. We used a touchdown program of 90°C for 3 min, 20 cycles of: 90°C for 30s, 60°C for 30s, and 72°C for 30s, 10 cycles of: 90°C for 30s, 50°C for 30s, and 72°C for 30s, and an additional 72°C for 10s, and cooled to 4°C. We precipitated PCR products for 15 min at room temperature with 64 µl of 100% ethanol and 16 µl Milli-Q $H_2O$, then centrifuged at 2000 xg for 30 mins in an Eppendorf 5804 centrifuge, and treated with 150 µl of 70% ethanol and centrifuged at 700 xg for 1 minute. We treated each sample with 12 µl of formamide and 0.125 µl of ABI 400 HD Analysis or ABI GeneScan 500-Rox ladder depending on estimated product size. To prepare for genetic analysis, we centrifuged the samples at 300 xg for 1 min, heated them at 95°C for 5 min, and cooled them on ice for 5 min.

We analyzed microsatellite fragments using ABI Gene Scan and Genotyper after running on an ABI Prism ® 3100 Genetic Analyzer (Applied Biosystems, Inc). We manually binned fragment lengths into

multiples of three basepairs, since these were triplet repeats, and small variations in length were unlikely to be biologically meaningful. We used triplet repeat number in all subsequent analyses. The microsatellites were quite variable and on averaged had about 8 alleles.

Linkage disequilibrium:

We used D' to measure linkage disequilibrium (Lewontin 1964b). In theory, this measure is standardized so that it varies between zero and one, with zero being no linkage and one being total linkage. We employed two different correction schemes to D', one based on resampling and one based on the value of D' at other loci. Any clone with missing data at one locus was omitted from all pairwise calculations involving that locus.

We omitted all loci on chromosome 2 for analysis of physical distance because it has a large-scale duplication in the reference genome sequence that is known to be absent in wild clones.

Simulations:

To look at the effect that sample size and number of alleles at each locus had on the linkage disequilibrium measures we used, we ran simulations where we generated sample populations that were in linkage equilibrium. To look at how the number of alleles at each locus impacts the measure of linkage disequilibrium, we ran four different sets of simulations of haploid individuals with two loci, each with a different

number of alleles (2,4,6, and 8) at each locus. For each of these, we modeled the population so that alleles were present in equal frequencies and the population size was infinite. From this population, we generated a set of samples. The alleles were picked from the population at random. This process was repeated independently for the second locus, so that the two loci should be in equilibrium. We then discarded the sample population if one or both of the loci only had one allele. We repeated this, generating samples of every size from 2 individuals to 1000 individuals. For each sample size and allele number combination, we ran 100 replicates, not including the discarded samples. We calculated the minimum, maximum, and average for each LD statistic.

LD value corrections:

We used resampling without replacement to correct for bias in D' following Devlin *et al.* (Devlin et al. 2001). We resampled the alleles without replacement from each of the loci and reassigned them randomly. This preserves allele number and frequency and while generating a set of haplotypes that should be in linkage equilibrium. We then used this set of haplotypes to estimate $D'_r$. We shuffled this dataset 100 times. The linkage between the two different loci is broken up while preserving sample size, allele number, and allele frequency. We then subtracted the

average value of the resampled replicates from the actual value to get a corrected D', D' $_{(r)}$. $D'_{(r)} = D' - \frac{\sum D'_{resampled}}{n}$

We also used D' values at nonsyntenic loci to correct D' values for syntenic loci. We first calculated D' values for all locus pairs in the dataset. For every comparison where the two focal loci were on the same chromosome, we took all the D' values between either of the focal loci and a locus on another chromosome and averaged them. We then subtracted this average from the D' value, to get a corrected D', D' $_{(c)}$.

If D'$_{(f1,f2)}$, D'$_{(f1,other)}$ and D'$_{(other,f2)}$ represent the D' values between the two focal loci, and the focal loci and loci on other chromosomes respectively.

$$D'_{(c)} = D'_{(f1,f2)} - \frac{\sum_{l_{other}} D'_{(f1,other)} + \sum_{l_{other}} D'_{(other,f2)}}{n}$$

In this paper we calculate D' $_{(c)}$ only for focal loci on the same chromosome, but the estimate can also be applied to loci on different chromosomes. With the "other" loci differing in the two summations.

**Results:**

To look for evidence of a history of sex in *D. discoideum*, we compared the D' values of adjacent pairs of loci, with those between more distant loci on the same chromosome and between loci on different

chromosomes. Crossing over should occur more frequently between two loci that are more distant than between loci that are close together. Even in the absence of crossing over, segregation reduces linkage disequilibrium between loci on different chromosomes. On average, D' between loci on different chromosomes was lower than differences between loci on the same chromosome and less than 0.1 mb apart (Figure 3.1, Mann-Whitney U test, $p>0.0001$). However the uncorrected D' values were uniformly very high (average 0.803, stdev 0.173). Because our dataset has a small number of individuals (though the same as the SNP study) and high numbers of alleles, the high LD may be due to bias.

To investigate the effects of sample size and allele number on bias in D' values, we then generated simulated datasets with no linkage disequilibrium and varying sample sizes and number of alleles. According to our simulations, both sample size and the number of alleles at each locus play an important role in determining what the bias of D' is (Figure 3.2). When small sample sizes are combined with a large number of alleles at each locus, the values of D' at completely unlinked loci can be quite high, though there is some bias even with large sample sizes and few alleles. With our study of *D. discoideum*, we had 24 individuals and an average of 8 alleles per locus. In our simulations, samples of unlinked loci with 8 (equally frequent) alleles per locus and 24 individuals had an

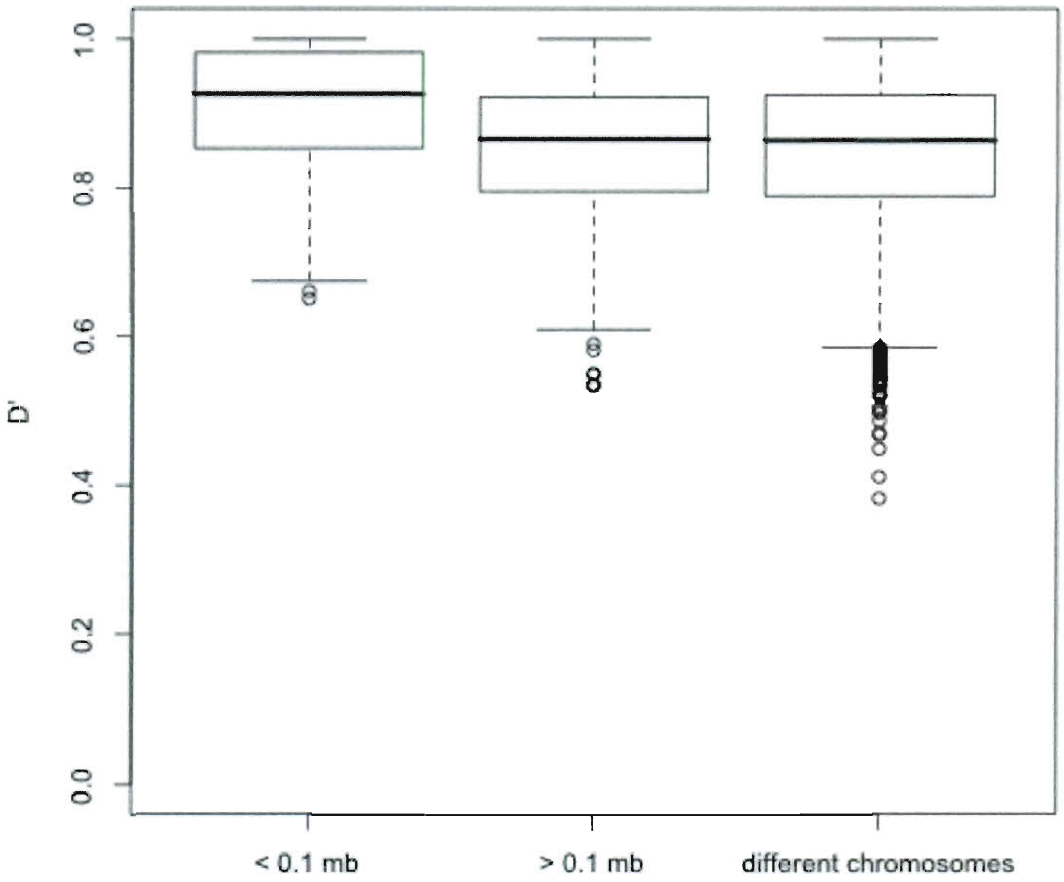



**Figure 3.1** D' values of loci less than 0.1 mb apart compared with D' values from more distant loci from on the same chromosome and D' values between loci on different chromosomes. A. Uncorrected D' measures (Mann-Whitney U test between close loci and loci on different chromosomes: $p<0.0001$, between close loci and distant loci: $p<0.0001$).

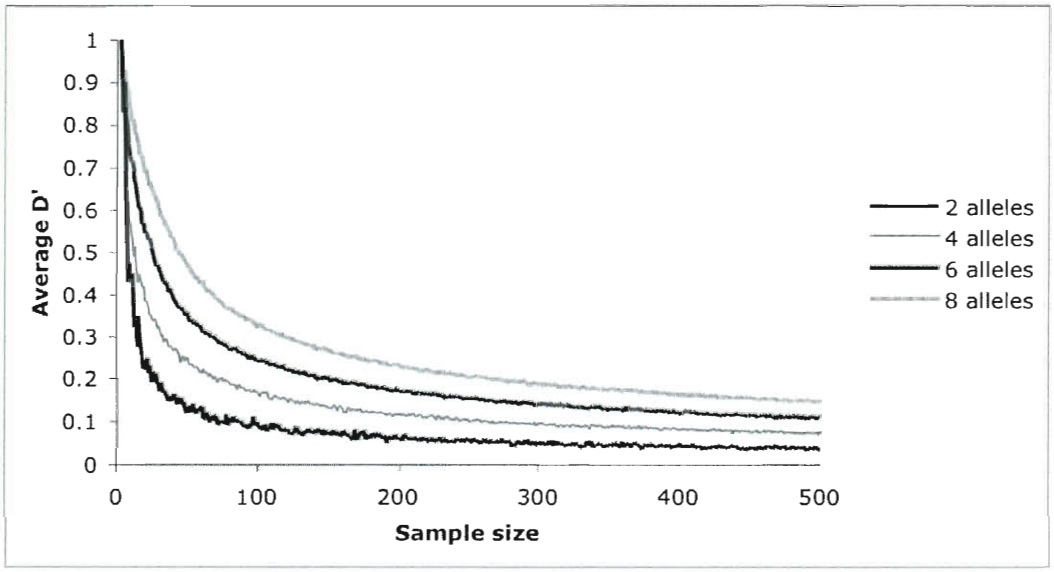



**Figure 3.2** Changes in D' as sample size and allele number change in a randomly generated dataset. Individual haplotypes are generated randomly, with no linkage disequilibrium. The average LD value is generated from 1000 runs at each sample size. We ran simulations for loci with 2, 4, 6 and 8 alleles at each locus.

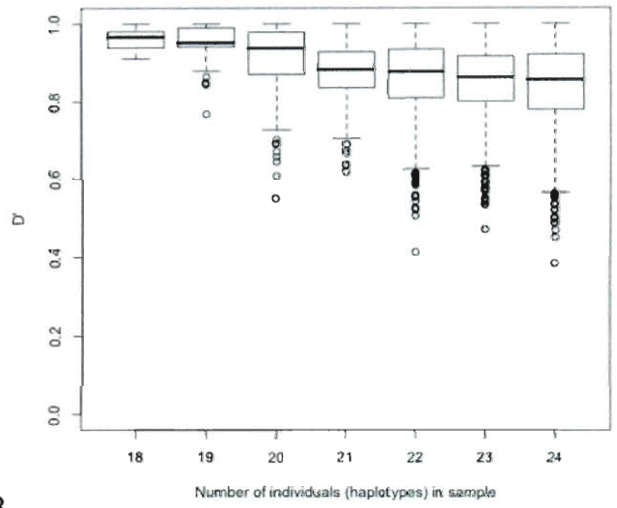
average D' of 0.6804, so bias explains the major part of why D' was so high.

To examine the bias that multiple alleles have on D' in our dataset, we plotted uncorrected D' as functions of the number of alleles and the sample size (Figure 3.3A and B respectively). The patterns are as expected from bias. D' tends to decrease with increasing sample size and increase with an increasing number of alleles (sample size: slope of linear regression: -0.013442, $p<0.0001$, number of possible haplotypes: slope of linear regression: 0.0015273, $p<0.0001$).

To attempt to correct this bias, we employed two different correction schemes, one based on resampling without replacement (Devlin et al. 2001) $D'_{(r)}$, and one based on D' values with loci on other chromosomes, $D'_{(c)}$. Both corrections decreased the effect of sample size (Fig. 3.3 C,E; $D'_{(r)}$ slope of linear regression: -0.0051313, $p<0.0001$, $D'_{(c)}$ slope of linear regression: 0.0031743, $p<0.001$). Both corrections decreased the effect of the number of possible haplotypes (Fig. 3.3 D,F; $D'_{(r)}$ slope of linear regression: 5.342e-4, $p<0.0001$, $D'_{(c)}$ slope of linear regression: -5.171e-4, $p<0.0001$).

The effect of sex can also be examined by how D' changes with physical distance. Uncorrected D' values were uniformly high, regardless of distance across chromosomes (Figure 3.4 A.). The resampling

A




B

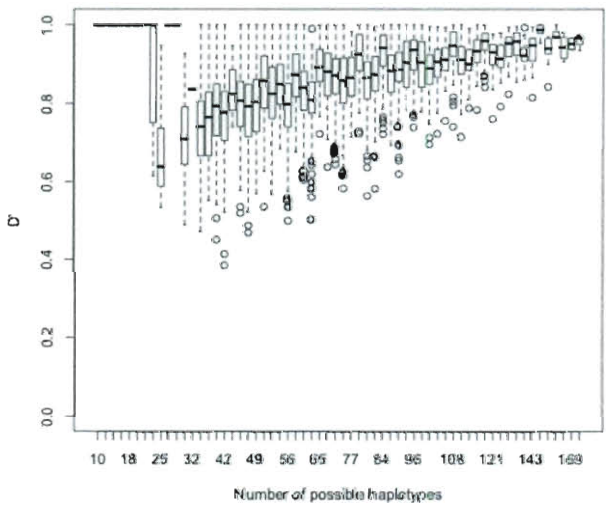
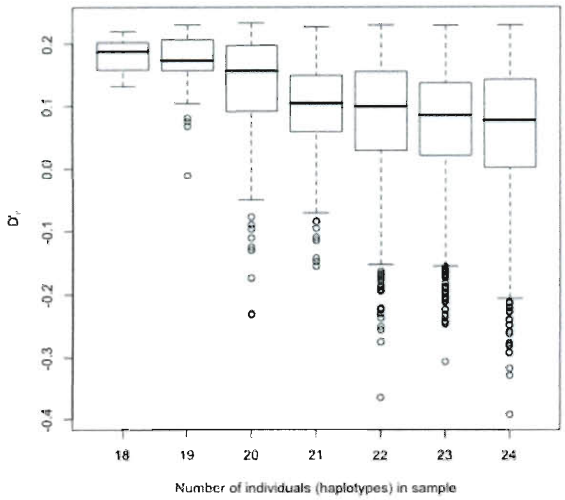



**Figure 3.3** Changes in D' as sample size (haploid individuals) and haplotype number change in *D. discoideum* microsatellite dataset. A. Changes in D' with sample size. (slope of linear regression = -0.013442, $p<0.0001$). B. Changes in D' with number of possible haplotypes. (slope of the regression = 0.0015273, $p<0.0001$)

C




D

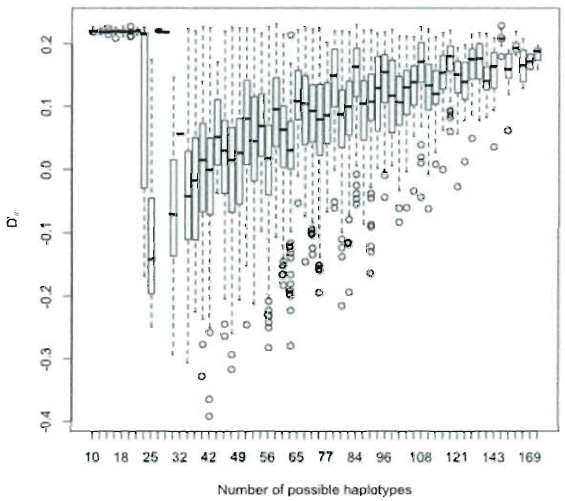
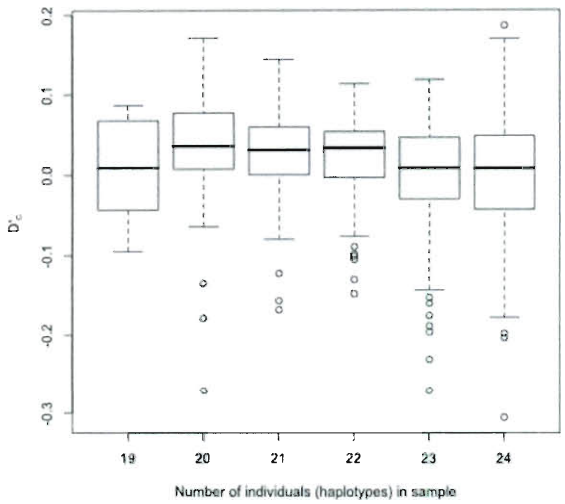



**Figure 3.3** C. Changes in $D'_{(r)}$ with the sample size (slope of linear regression = -0.013496, $p<0.0001$). D. Changes in $D'_{(r)}$ with the number of possible haplotypes (slope of linear regression = 5.342e-04, $p<0.0001$).

E




F

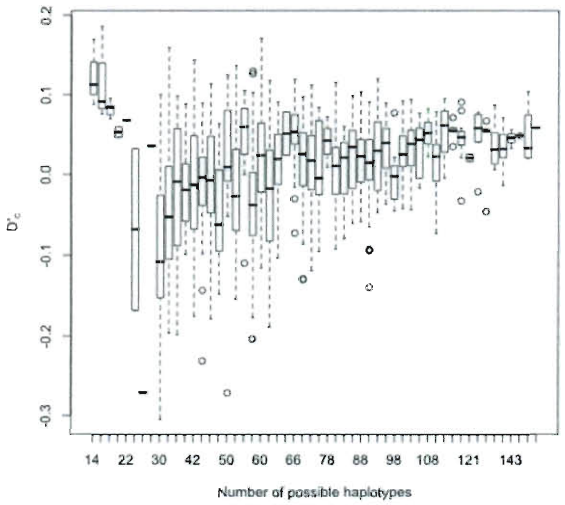
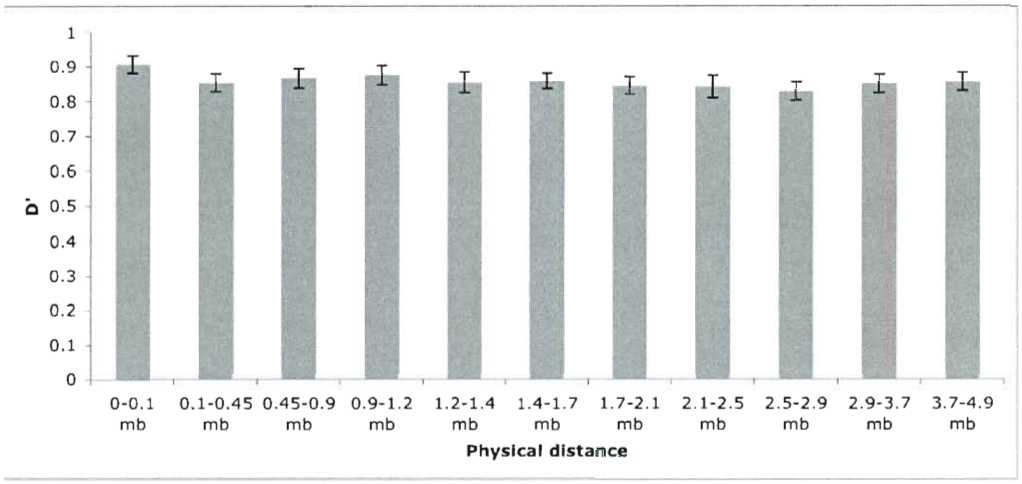



**Figure 3.3** E. Changes in $D'_{(c)}$ with the sample size (slope of linear regression = -0.007948, $p<0.001$). F. Changes in $D'_{(c)}$ with the number of possible haplotypes (slope of linear regression = 6.134e-04, $p<0.0001$).

A




B

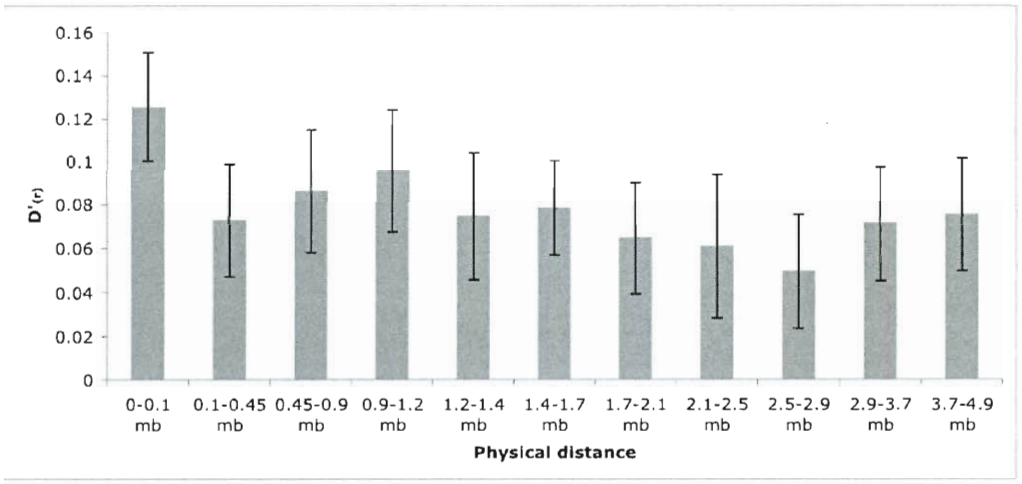



**Figure 3.4** *D. discoideum* syntenic D' values plotted against the physical distance between the two loci. All pairs on chromosome 2 have been omitted (see methods for details). Bins on the histogram have been chosen so that the sample size remains relatively constant over bins. Error bars are 95% confidence intervals. A. Uncorrected D' values B. $D'_{(c)}$ values

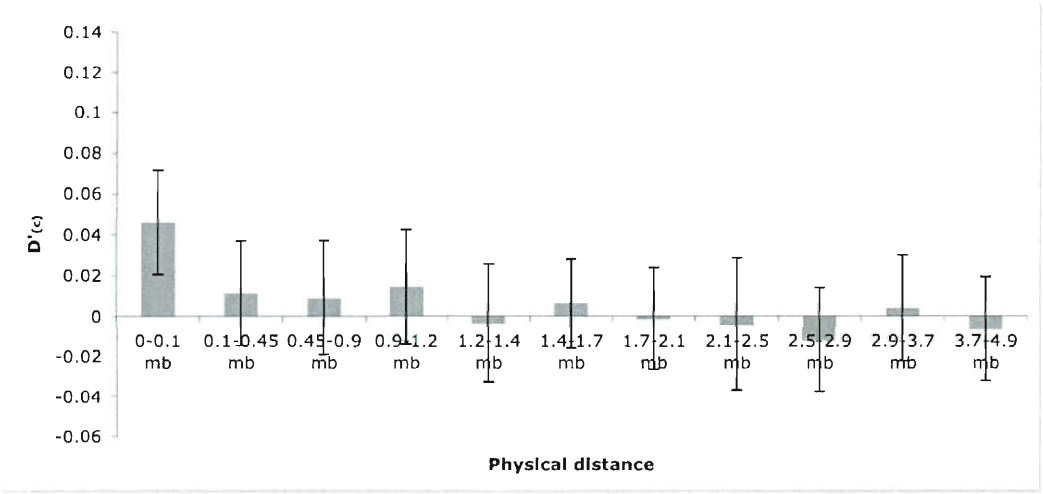



**Figure 3.4** continued. C. $D'_{(r)}$ values.

correction, $D'_{(r)}$, decreased values across all distances (Figure 3.4 B). Loci that were between 0 to 0.1 mb had the highest $D'_{(r)}$ and $D'_{(r)}$ was uniformly lower at farther distances (Mann-Whitney U test between close loci and loci on different chromosomes: $p<0.0001$, between close loci and distant loci: $p<0.0001$). Using our correction, $D'_{(c)}$, linkage disequilibrium decays rapidly as the physical distance between loci increases (Figure 3.4 C.). Increased $D'_{(c)}$ values occur when the two loci were less than 0.1 mb apart and decayed to levels similar to those between nonsytenic loci at further distances (Figure 3.4 C, Mann-Whitney U test between close loci and distant loci: $p<0.0001$).

**Discussion**:

We looked at linkage disequilibrium in *D. discoideum* using microsatellites to calculate D'. We found uniformly high D' values, which contrasted with other recent work estimating linkage disequilibrium in *D. discoideum* using less variable SNPs and other microstatellites (Landi 2003; Flowers et al. 2010). Both corrections we looked at, using resampling and using non-syntenic loci decreased the value of D'. After these corrections, our results were consistent with Flowers et al. (2010).

Many studies have avoided using microsatellites for linkage disequilibrium because of the bias inherent in multiallelic datasets (Ardlie

et al. 2002). However, using biallelic data, such as SNPs, can result in bias as well (Figure 3.3). In both SNPs and microsatellite markers, larger sample sizes still show some bias. This implies that corrections should be useful, even when sample sizes are in the hundreds.

In order to control this bias, we used two different correction schemes, one based on resampling (Devlin et al. 2001) and one based on the D' values generated at non-syntenic loci. Both of these corrections greatly reduced the bias due to sample size and allele number, with $D'_{(c)}$ reducing values closer to zero than $D'_{(r)}$ We compared the D' of the focal loci to the average D' of all other loci because a large part of the variation in D' values stems from allele number, sample size and frequencies of alleles. This method reduced bias due to sample size more than using a resampling correction did. This method is less computationally intense, but requires information about markers throughout the genome. This may be ideal for whole genome datasets, such as the recent population genomics study of seven inbred lines of *Drosophila simulans* (Begun et al. 2007), where there is in depth coverage of just a few individuals.

This study documents the changes that occur in D' as sample size and allele number change. In theory, D' should range from 0 to 1, with 0 being complete equilibrium (Lewontin 1988). However, our simulations show that minimum D' for unassociated alleles increases as the number of

alleles increase and as the sample size decreases. This makes comparing D' values difficult if the data involved do not have very similar characteristics.

Our results confirm that linkage disequilibrium decreases rapidly over the length of the *D. discoideum* chromosome, as is the case with SNPs (Flowers et al. 2010). This suggests that laboratory studies where genetic recombination during the macrocyst cycle was not observed (Lamphier and Yanagisawa 1983; Francis and Eisenberg 1993) do not reflect the situation in natural settings. Recombination does occur.

This study improves the status of *D. discoideum* as a model organism by adding to the knowledge of basic biology of this organism. This has implications for research on the social interactions of *D. discoideum*. Recombination serves to mix up different parts of the genome meaning that relatedness of two individuals at one locus could be completely independent of relatedness of other loci or chromosomes. *D. discoideum* has sex in natural populations; this suggests that efforts to understand the conditions under which macrocysts hatch are not in vain. Understanding the optimum conditions for macrocyst hatching will provide a new avenue for genetic studies to construct new strains.

In looking at the basic question of whether recombination occurs, both microsatellites and SNPs give the same result. However, because

microsatellites have a higher mutation rate than the rest of DNA (Ellegren 2000), they may be better at detecting recent disequilibrium events, such as separation of populations. Microsatellites may be the better marker to use when markers are farther apart, resulting in fewer markers to cover the genome (Varilo et al. 2003; Mueller 2004). Studies using microsatellites should employ a correction factor to ensure that results are not an artifact of setup of the dataset. Comparisons between uncorrected values should be avoided, not just between datasets but also within the same dataset. Because good corrections are available, microsatellites need not be avoided in studies of linkage disequilibrium.

**References:**

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S et al. (2001) Extent and distribution of linkage disequilibrium in three genomic regions. American Journal of Human Genetics 68(1): 191-197.

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3(4): 299-309.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP et al. (2007) Population genomics: Whole-genome analysis of polymorphism

and divergence in Drosophila simulans. Plos Biology 5(11): 2534-2559.

Chbel F, Broderick D, Idaghdour Y, Korrida A, McCormick P (2002) Characterization of 22 microsatellites loci from the endangered Houbara bustard (*Chlamydotis undulata undulata*). Molecular Ecology Notes 2(4): 484-487.

Devlin B, Roeder K, Otto C, Tiobech S, Byerley W (2001) Genome-wide distribution of linkage disequilibrium in the population of Palau and its implications for gene flow in Remote Oceania. Human Genetics 108(6): 521-528.

Devreotes P (1989) Dictyostelium discoideum - a model system of cell-cell interactions in development. Science 245(4922): 1054-1058.

Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. Trends in Genetics 16(12): 551-558.

Flowers J, Li S, Stathos A, Saxer G, Ostrowski E et al. (2010) Variation, sex, and social cooperation: molecular population genetics of the social amoeba *Dictyostelium discoideum*. Plos Genetics 6(7).

Francis D (1998) High frequency recombination during the sexual cycle of *Dictyostelium discoideum*. Genetics 148(4): 1829-1832.

Francis D, Eisenberg R (1993) Genetic structure of a natural population of *Dictyostelium discoideum*, a cellular slime mold. Molecular Ecology 2(6): 385-391.

Hedrick PW (1987) Gametic disequilibrium measures - proceed with caution. Genetics 117(2): 331-341.

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38(6): 226-231.

Hudson RR, Kaplan NL (1985) Statisical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111(1): 147-164.

Kessin RH (2001) Dictyostelium: evolution, cell biology, and the development of multicellularity. Developmental and Cell Biology Series 38: i-xiv, 1-294.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics 22(2): 139-144.

Lamphier MS, Yanagisawa K (1983) Induction of macrocyst formation by factors secreted by giant cells in *Dictyostelium discoideum*. Development Growth & Differentiation 25(5): 495-501.

Landi M (2003) Reproductive conflicts in the social wasp, *Eustenogaster fraterna*, and in the social amoeba, *Dictyostelium discoideum*. Houston, TX: Rice University.

Lewontin RC (1964a) Interaction of selection and linkage. 1. General considerations - heterotic models. Genetics 49(1): 49-&.

Lewontin RC (1964b) Interaction of selection and linkage. 2. Optimum models. Genetics 50(4): 757-&.

Lewontin RC (1988) On measures of gametic disequilibrium. Genetics 120(3): 849-852.

McRae AF, McEwan JC, Dodds KG, Wilson T, Crawford AM et al. (2002) Linkage disequilibrium in domestic sheep. Genetics 160(3): 1113-1122.

Mueller JC (2004) Linkage disequilibrium for different scales and applications. Briefings in Bioinformatics 5(4): 355-364.

Nickerson AW, Raper KB (1973) Macrocysts in life cycle of dictyosteliaceae. 2. Germination of macrocysts. American Journal of Botany 60(3): 247-254.

Raper KB (1984) The dictyostelids. The dictyostelids: i-xi, 1-453.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.

Schurko AM, Neiman M, Logsdon JM (2009) Signs of sex: what we know and how we know it. Trends in Ecology & Evolution 24(4): 208-217.

Slate J, Pemberton JM (2007) Admixture and patterns of linkage disequilibrium in a free-living vertebrate population. Journal of Evolutionary Biology 20(4): 1415-1427.

Slatkin M (2008) Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9(6): 477-485.

Strassmann JE, Zhu Y, Queller DC (2000) Altruism and social cheating in the social amoeba Dictyostelium discoideum. Nature 408(6815): 965-967.

Sussman M (1966) Biochemical and genetic methods in the study of cellular slime mold development. In: Prescott D, editor. Methods in Cell Physiology. NY: Academic Press. pp. 397-410.

Varilo T, Paunio T, Parker A, Perola M, Meyer J et al. (2003) The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. Human Molecular Genetics 12(1): 51-59.

Weir BS, Cockerham CC (1978) Testing hypotheses about linkage disequilibrium with multiple alleles. Genetics 88(3): 633-642.

Williams RSB, Boeckeler K, Graf R, Muller-Taubenberger A, Li ZR et al. (2006) Towards a molecular understanding of human diseases using Dictyostelium discoideum. Trends in Molecular Medicine 12(9): 415-424.

Zapata C, Carollo C, Rodriguez S (2001) Sampling variance and distribution of the D' measure of overall gametic disequilibrium between multiallelic loci. Annals of Human Genetics 65: 395-406.

Zaykin DV, Pudovkin A, Weir BS (2008) Correlation-based inference for linkage disequilibrium with multiple alleles. Genetics 180(1): 533-545.

# CHAPTER 4

## Nucleotide content and selection pressure in two GC-poor organisms, *Dictyostelium discoideum* and *Plasmodium falciparum*

**Abstract**:

Several processes act to influence the nucleotide content of sequences, but how these processes interact to influence nucleotide content is unclear. Diminished GC content could be the result of non-adaptive forces such as mutation and biased gene conversion. Selection could also change nucleotide content by either systematically removing variation that changes nucleotide content or systematically selecting for a specific nucleotide content. The nucleotide content of a genome should be a balance between several processes including mutation and selection. Because selection can only act upon the variation that is generated by mutation, if mutation is biased towards generating AT from GC, purifying selection should often act to remove that variation, and maintain the original GC content. Thus, sequences that are under more purifying selection should have lower GC content than sequences that are not under purifying selection. To investigate this, we used the genomes of two GC-poor organisms, *Dictyostelium discoideum* and *Plasmodium falciparum* to investigate whether purifying selection influences nucleotide content. We compared GC contents of sequences predicted to be under greater or less purifying selection. The classes that we compared included coding regions with untranslated sequences, the first and second codon position as compared to the third codon position, domains with non-

domains in coding regions, genes with pseudogenes, and looking at expression level and nucleotide content. In all classes that we examined, GC content was lower in regions expected to be under more purifying selection. This shows that the balance between mutational bias and purifying selection plays a major role in shaping the nucleotide content of sequences. It also suggests that nucleotide content could serve as an index of purifying selection, at least in genomes with strong mutation bias.

**Introduction**:

Genomes vary in GC content because of the historical interplay between mutation rates, recombination and selection (Pozzoli et al. 2008; Rocha and Feil 2010). Studies of mutation rates reveal that GC content is not determined by mutation rate alone (Marais 2003; Duret 2006; Lind and Andersson 2008; Lynch et al. 2008; Hershberg and Petrov 2010; Hildebrand et al. 2010). Numerous direct explanations for the nucleotide GC content have been proposed (Rocha and Feil 2010). For example, GC rich sequences are more thermostabile, so thermophiles might be expected to have higher GC contents (Kagawa et al. 1984). Similarly there might be selection against AT-rich dinucleotides, as they are more susceptible to dimerization with UV radiation (Singer and Ames 1970).

However, genomic GC composition influences the composition of

coding sequences rather than the composition of coding sequences determining the GC content (Knight et al. 2001). If GC content was the result of amino acid or codon usage, then there would be many ways to generate sequences of any given GC content. Different organisms that have the same GC content would not be expected to have the same proportions of amino acids. However, unrelated organisms that have similar GC contents also have similar amino acid and codon usage. This suggests that the amino acid compositions are a result of the GC content of the genome rather than the other way around. (Knight et al. 2001).

However, mutations are not the sole factor in determing GC content. Mutations are universally biased towards generating AT and removing GC from sequences (Lynch et al. 2008; Hershberg and Petrov 2010; Hildebrand et al. 2010). If mutations were the sole factor in determine GC content, then GC contents would be uniformly low, and sequences would reach an equilibrium state. Here the number of mutations to A or T would equal those to C or G, where the increased rate of GC to AT mutations is compensated for by the decreased frequency of GC. However, in both bacteria and yeast, there is still an excess of GC to AT mutations (Lynch et al. 2008; Hershberg and Petrov 2010; Hildebrand et al. 2010). This discrepancy in GC content is usually attributed to selection maintaining more GC-rich sequences (Rocha and Feil 2010).

While selection acts to produce proteins with a higher fitness, it can only act on the variation generated by mutation and recombination. Mutations are typically biased towards generating A/T from G/C. A common source of mutations is the deamination of cytosine, promoting C->T transitions (Shen et al. 1994). These mutations frequently are initiated in single-stranded DNA. The rate at which CpG mutates to T actually increases in regions of low GC content because high AT regions are more prone to becoming single stranded (Fryxell and Moon 2005; Morton et al. 2006).

Under the nearly neutral mutation theory (Ohta 1992), most mutations are neutral or very nearly so. This suggests that many mutations including those that change amino acids will be of little effect. Several studies on changes to genomic nucleotide content have shown that as GC content varies so do the amino acids that are used (review: (Gautier 2000).

Purifying selection will act to remove many of the mutations that appear in genes, particularly those in regions that are functionally constrained. Regions that are more functionally constrained should show a higher GC content because fewer mutations persist in these regions. Mutations, which are generally biased towards generating AT from GC will be removed under purifying selection, preserving the original nucleotide content. Sequences with less purifying selection should show a decrease

in GC content as mutations occur and persist in the sequence.

GC content could therefore serve as an indicator of the relative amount of purifying selection that a sequence is under. We test this idea by comparing the GC content of sequences that are under different levels of purifying selection in two GC-poor organisms, *Dictyostelium discoideum* and *Plasmodium falciparum*. We chose these two organisms because of their low GC contents. These organisms probably have strong mutational bias towards AT. If GC content is correlated with the amount of purifying selection, then GC content could serves as an indicator for purifying selection.

This measure would complement rather than supplant traditional comparative measures of selection, such as dN/dS. All such measures are imperfect. For example, there is sometimes selection on the synonymous sites of proteins, which are used to calculate dS (Ophir et al. 1999; Jiang and Govers 2006). This can lead to a miscalculation of the amount of purifying selection on a sequence (Ophir et al. 1999). An additional measure of purifying selection, such as GC content could help correct this problem. In addition, because CG content can be measured on a single genome, it provides a measure that does not require comparative data from other genomes.

Plotkin, Dushoff and Fraser attempted to use the codon sequence

of a single sequence to generate a measure of directional selection (Plotkin et al. 2004). Their codon volatility index attempts to measure the proportion of potential mutations that are synonymous out of all possible mutations as a measure of the strength of selection. Several criticisms of their measure have been made (Chen et al. 2005; Hahn et al. 2005; Nielsen and Hubisz 2005). Our measure has a different goal; rather than attempting to predict the amount of directional selection present using a single sequence, we investigated the nucleotide content of different classes of sequences that should be under different amounts of purifying selection.

If we assume that GC to AT mutation rates are equal in a genome, then sequences with more purifying selection will resist this pressure towards AT more because mutations are selected against. This leads to the prediction that sequences which under more purifying selection will have higher GC contents. This prediction can be tested by comparing several classes of sequences that are under differing amounts of purifying selection. First, non-coding (intragenic) regions should have a lower average GC content than coding regions (Graur and Li 2000). Second, the largely synonymous 3rd codon position should have a lower average GC content than the 2nd and the 1st codon positions because changing the 3$^{rd}$ codon position does not always change the amino acid sequence

of the protein. Non-domain coding regions should have a lower average GC content than protein domains because domains are the functional units of proteins (Koonin et al. 2002). Pseudogenes should have a lower average GC content than their functional ancestors. Essential genes should have a lower average GC content than non-essential genes.

We have tested these predictions on a genome-wide scale in the GC poor eukaryotes, *D. discoideum* and *P. falciparum*. If these hold true, then GC content can serve, at least to some degree, as an indicator of purifying selection and therefore perhaps a rough indicator of the importance of genes and other sequences.

*P. falciparum* is an unicellular eukaryote that is the causative agent of malaria in humans. It is haploid except the diploid stages immediately following sex, with a genome of 22.8 mb over 14 chromosomes. It has a GC content of roughly 19.4%, making it one of the lowest GC genomes sequenced (Gardner et al. 2002). *Dictyostelium discoideum* is a unicellular eukaryote that has long been used as a model organism in development (Devreotes 1989) and social behaviour (Strassmann et al. 2000). It is haploid, with a genome of 34 mb over 6 chromosomes (Eichinger et al. 2005).

Both of these genomes have a very low GC content. The overall GC content of *D. discoideum* is roughly 22.5% (Eichinger et al. 2005)

while the over all GC content of *P. falciparum* is roughly 19.4% (Gardner et al. 2002). We choose these GC-poor organisms because presumably mutational bias towards A/T from G/C influences the GC nucleotide content of the genome. If this mutational bias is stronger in these GC-poor organisms, we expect that the conflict between mutational bias and selection pressure should be strongest in these organisms. To look at how purifying selection and mutation interact, we split both of these genomes into classes of sequences that are thought to be under differing levels of selection, and compared the GC content between them.

**Methods**:

We downloaded whole genome fasta and corresponding genomic flat files (.gff file) from their respective database site (http://www.dictybase.org Dec 26, 2008, http://plasmodb.org version 5.5). Gene sequences were identified using the gff entries to pull sequences from the fasta file. We checked all genes to make sure that they started with ATG and ended with a stop codon. Introns were analyzed by identifying gaps between exons using the .gff file. We downloaded the InterPro domain information from the respective genome database sites. We used the predictions from Gene3D which uses the protein structure to predict domains. We did this in order to minimize the effects of GC

content on domain prediction. Upstream and downstream untranslated regions (UTRs) were generated by taking the adjacent sequence up to the next gene on the chromosome or 1 kb, whichever was shorter. In *P. falciparum,* pseudogenes were identified as those genes with the word 'pseudogene' in the gene description. In *D. discoideum*, Olsen (Olsen 2005) had previously identified a set of pseudogenes. Olsen used BLAST to search all identified genes against all untranslated regions, looking for matches with premature stop codons or frameshift mutations. Any sequence that was identified as a pseudogene was removed from all analyses involving coding sequences.

We calculated GC content for each category, including both overall and position based GC content where applicable. We downloaded EST sequence libraries from dictybase. The libraries were generated from cells at different stages in the *D. discoideum* lifecycle (Muramoto et al. 2003; Urushihara et al. 2004). ESTs were blasted against the genomic sequence for *D. discoideum*. Hits were filtered so that only the longest total hit region (covering >= 75% of the EST query). Out of these, we filtered out ESTs that did not overlap with any gene predictions. We filtered out all genes that had no EST matches in any of the libraries. We counted the number of ESTs from each developmental stage as expression level at that stage, grouping libraries generated from the same

stage. Though gene expression has been analysed in *P. falciparum*, we are unclear on the relative importance of the different lifecycle stages. Because of this, we have avoided analyzing expression in *P. falciparum*.

In our analyses, we used 12,644 genes from *D. discoideum* and 5,464 genes from *P. falciparum*. For the pseudogene analysis, we had 1,459 pseudogenes in *D. discoideum* and 69 pseudogenes in *P. falciparum*. Pseudogenes were identified as discussed above. A total of 4,177 genes were included in the expression analysis of *D.discoideum* genes. The list of protein domains from the InterPro database was downloaded from dictybase.org (on May 8, 2008). This list was used to split codons into two different categories, those coding for amino acids in domains, and those coding for amino acids not associated with any domain annotation.

We used R to calculate all statistics. We used paired t-tests where comparisons were being made between different parts of the same locus to control for any possible variation in background GC content and unpaired t-tests elsewhere.

**Results and Discussion:**

To look at how mutation and selection interact, we broke the genome into several classes that are thought to be under differing levels

of selection, and compared the GC content between them. First we examined the coding regions vs the non-coding regions. GC content is higher in coding regions than non-coding regions in both *D. discoideum* and *P. falciparum* (Figure 4.1, paired t-test between genic region and 1 kb up/down-stream $p<0.001$ both organisms). Within genes, GC content was lower in introns than in exons in both *D. discoideum* and *P. falciparum* (Figure 4.2; *D. discoideum* exons 28.4%, introns 10.7%; *P. falciparum* exons 25.9%, introns 13.5%, paired t-test, $p<0.001$ both species). With the exception of splice site junctions, the sequence of introns is less important than that of the protein coding exons. Introns have remarkably low GC contents, which may be due, in part, to selection on introns to be distinguishable from exons (Schwartz et al. 2009). Lower GC content in introns is consistent with our explanation that introns have experience less purifying selection, and thus are less able to resist a mutational bias towards AT, especially in *P. falciparum*, where the GC content of introns is similar to the GC content of other non coding regions.

In *D. discoideum* and *P. falciparum*, GC content was lowest in the third codon position (Figure 4.3; *D. discoideum* 16.6%, paired t-test between $1^{st}$ and $3^{rd}$ position $p<0.01$, between $2^{nd}$ and $3^{rd}$ position $p<0.01$; *P. falciparum* 18.5%, paired t-test between $1^{st}$ and $3^{rd}$ position $p<0.01$, between $2^{nd}$ and $3^{rd}$ position $p<0.01$), suggesting that many synonymous

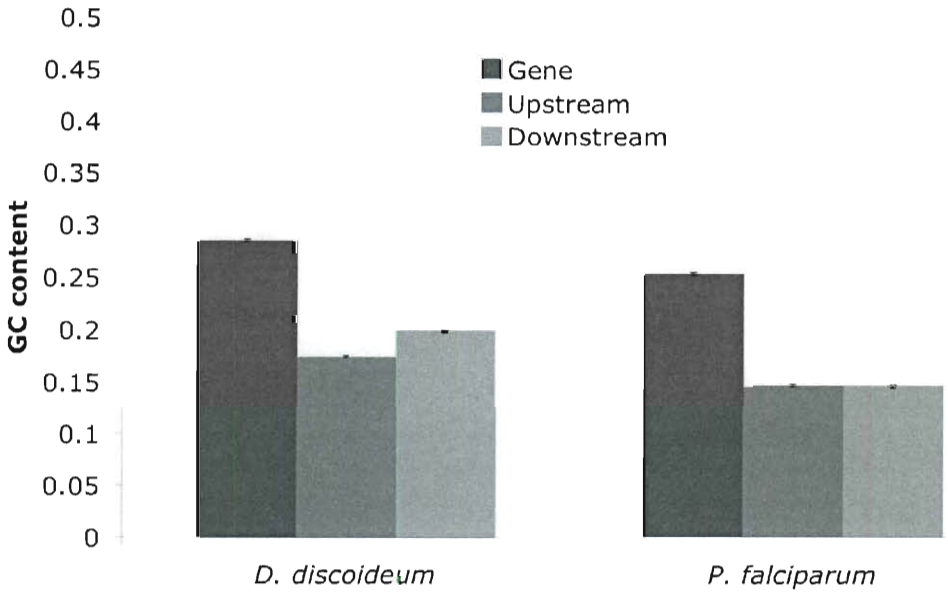



**Figure 4.1** GC content of genes compared to adjacent upstream and downstream untranslated regions in both *D. discoideum* and *P. falciparum*. In both organisms, there was no significant difference between the upstream and the downstream regions and, GC content of the coding regions of genes was significantly higher than either the upstream and the downstream regions. (Paired t-test, $p<0.001$ both cases. D. *discoideum* n=12,644. *P. falciparum* n = 5,464)

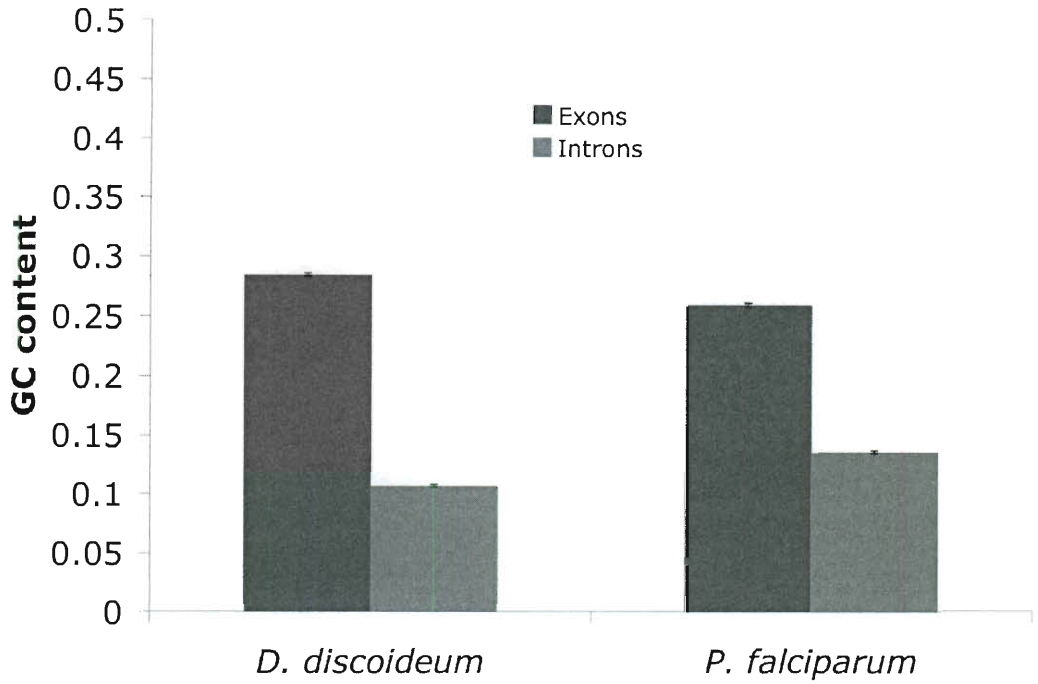



**Figure 4.2** GC content of introns compared to exons in both *D. discoideum* and *P. falciparum*. In both organisms, the GC content of exons was higher than that of introns. (Paired t-test, $p<0.001$ both cases. *D. discoideum* n= 12,644. *P. falciaprum* n=5,464.)

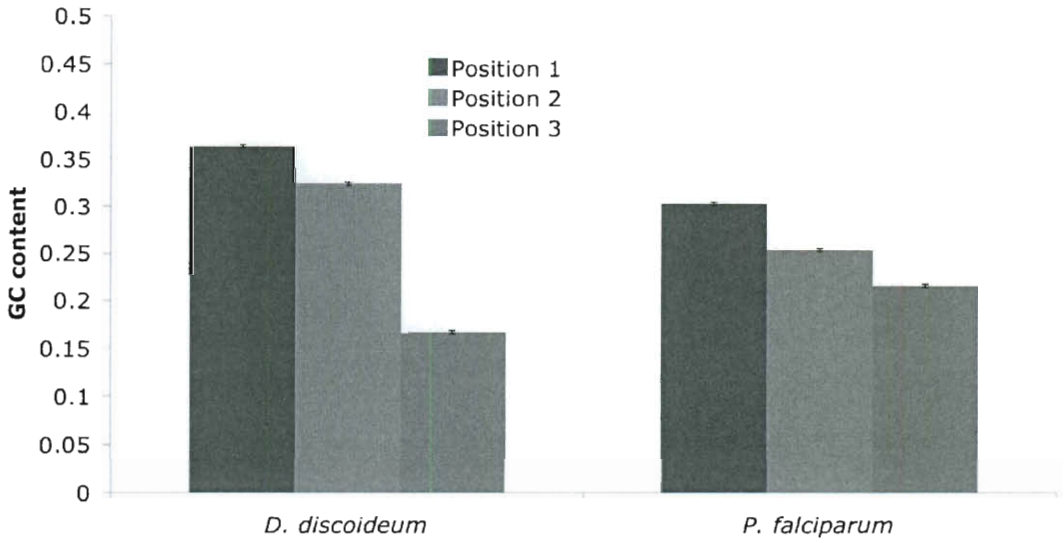



**Figure 4.3** GC content of each codon position in coding regions of genes in both *D. discoideum* and *P. falciparum*. In both organisms, the first two codon positions were significantly higher in GC content than the synonymous third codon position. (Paired t-test between GC1 and GC3 $p<<0.001$, between GC2 and GC3 $p<<0.001$ both species. *D. discoideum* n=12,644. *P. falciparum* n=5,464)

sites have experienced AT mutations. GC content was different between the first two codon positions as well (Figure 4.2; *D. discoideum* 1st position 36.4%, 2nd 32.4%, paired t-test, $p<0.05$; *P. falciparum* 1st position 33.0%, 2nd 24.3%, paired t-test, $p<0.05$). The difference in GC content between the first two and then third codon position is consistent with our expectations. The differences between the first two codon positions can also be explained. The second codon position influences the biochemistry of the resultant amino acid, with T associated with hydrophobic residues and A associated with hydrophilic residues (Pascal et al. 2006). The presence of G is increased at the first codon position in many organisms (D'Onofrio and Bernardi 1992). It has been hypothesized that G in this position increases translational efficiency (Gutierrez et al. 1996). Thus composition at the first and second codon positions will vary from each other and can be subject to different pressures.

Similarly to the comparison between intergenic regions and coding regions, GC content was lower in introns than in exons in both *D. discoideum* and *P. falciparum* (Figure 4.3; *D. discoideum* exons 28.4%, introns 10.7%; *P. falciparum* exons 25.9%, introns 13.5%, paired t-test, $p<0.001$ both species). With the exception of splice site junctions, the sequence of introns is less important than that of the protein coding exons. Introns have remarkably low GC contents, which may be due, in part, to

selection on introns to be distinguishable from exons (Schwartz et al. 2009). Lower GC content in introns is consistent with our explanation that introns have experience less purifying selection, and thus are less able to resist a mutational bias towards AT, especially in *P. falciparum*, where the GC content of introns is similar to the GC content of other non coding regions.

For *D. discoideum*, we have information on protein domains. Within coding sequences, those identified as known protein domains have a higher GC content than non-domain sequences. Domains had a higher GC content than the non-domain regions in the same gene (Figure 4.4, Overall: domains 29.3%, non-domains 28%, Position 1: domains 38%, non-domains 35.6%, Position 2: domains 33.7%, non-domains 31.9%, Position 3: domains 16.2%, non-domains 16.5% paired t-test, $p<0.005$ for all). This is consistent with the idea that protein domains are under more purifying selection than non-domain regions.

Pseudogenes have been identified in *D. discoideum* by searching the whole genome for coding regions that have been duplicated but have early termination codons (Olsen 2005). In *D. discoideum*, pseudogenes have a lower GC content than genes (Figure 4.5, pseudogenes 26.7%, genes 28.6%, t-test, $p<0.001$). An analysis of pseudogenes in several animal species shows that pseudogenes have an amino acid frequency

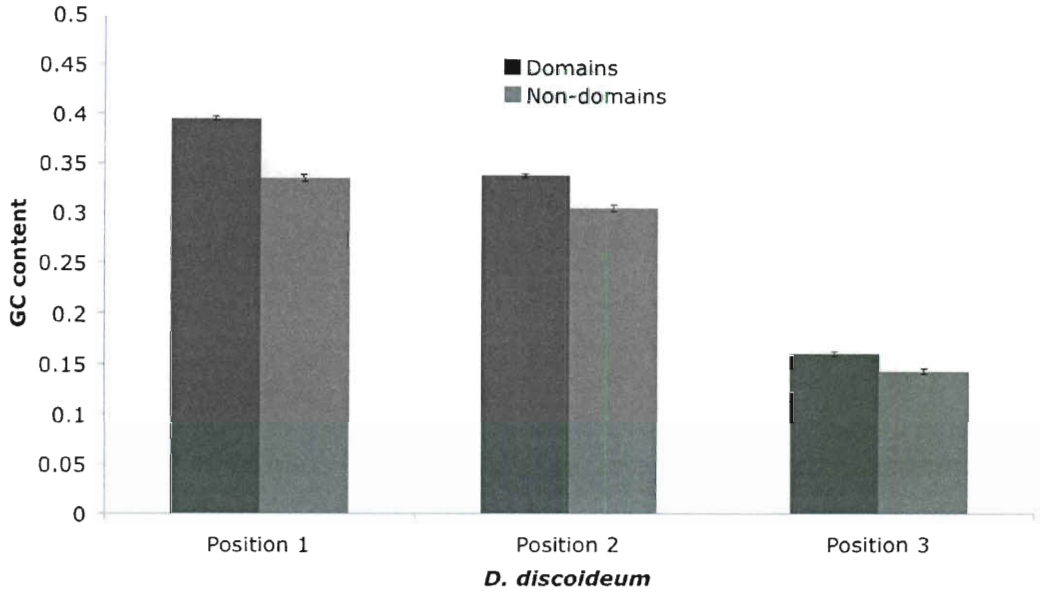



**Figure 4.4** GC content of DNA coding for domain and nondomain regions of proteins in *D. discoideum*. Domains were identified through the Interpro database, Gene3D. Domains had a higher GC content than non-domains both overall and in all three codon positions. (Paired t-test $p<0.001$ n= 4,376)

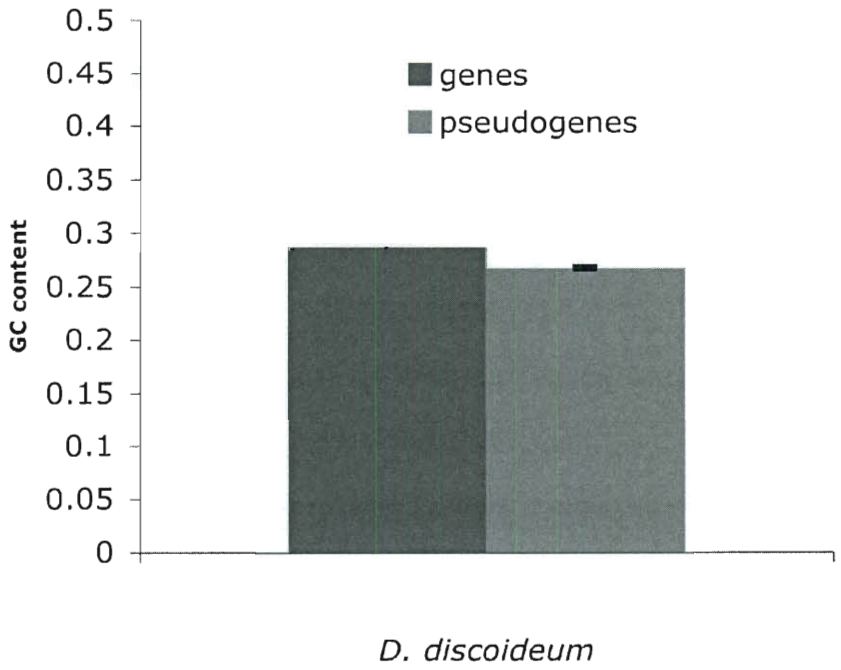



**Figure 4.5** GC content of genes and pseudogenes in *D. discoideum*. Genes have a significantly higher GC content than pseudogenes in *D. discoideum*. (t-test, $p < 0.001$ genes = 12,644 , pseudogenes = 1,458)

intermediate between coding sequences and artificially translated intergenic regions (Echols et al. 2002). This suggests that pseudogenes are no longer under purifying selection and are but are subject to mutational bias, like intragenic regions. Pseudogenes that are more divergent from their functional ancestors are expected to also have bigger differences in GC contents.

There is strong translational selection on highly expressed proteins to ensure that large amounts of the amino acid product are produced efficiently (Akashi 2001) and folded properly (Drummond and Wilke 2008). Part of this translation selection is the positive selection on the sequence to maintain optimal codon usage. A previous study using a subset of 47 genes in *D. discoideum* suggested that highly expressed genes were composed of more GC-rich codons than less expressed genes (Sharp and Devine 1989). In this study, they inferred relative amounts of gene expression by relying on the relative expression in other organisms, rather than measuring it directly. In our analysis, we use expressed sequence tags (EST) as a measure of mRNA abundance and gene expression. In *D. discoideum*, increasing GC content weakly correlates with an increasing number of EST tags in the vegetative stage (Figure 4.6. $R^2$ = 0.14). There has already been much previous research on gene expression in *P. falciparum*. As gene expression increases the usage of

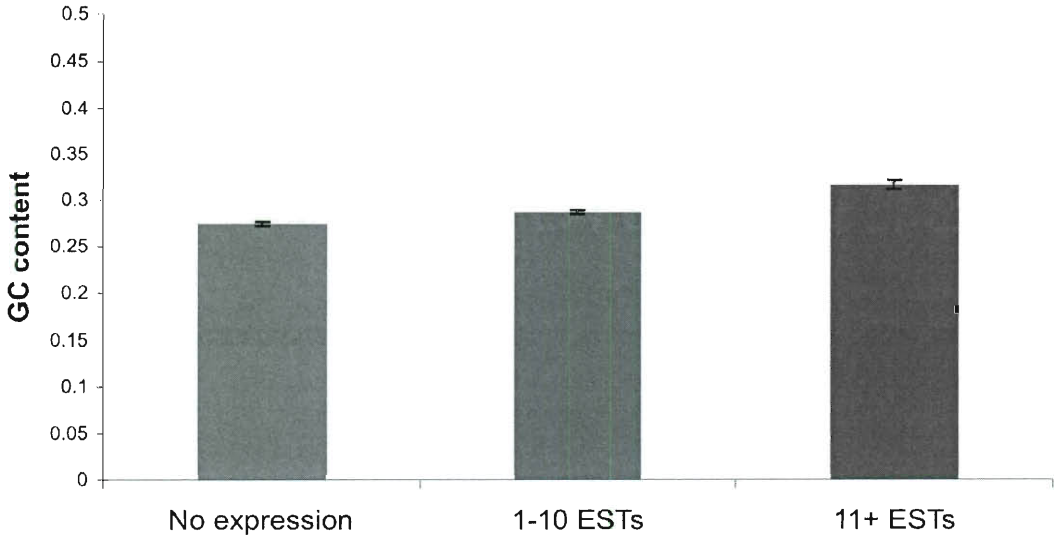



**Figure 4.6** GC content compared with expression level in *D. discoideum*. Expression level was measured as the number of EST hits present. See methods for further details. As expression level in the vegetative stage of *D. discoideum* increases, GC decreases slightly. ($R^2$= 0.14, $p$<<0.001 n=4,177.)

GC-rich synonymous codons increases (Chanda et al. 2005).

We have examined the GC content of classes of sequences that are under relatively different amounts of positive selection in two unrelated GC poor organisms. AT-biased mutations provide a force that is countered by purifying selection where it acts. Together, these two forces have a significant impact on nucleotide content. In every comparison that we made, the hypothesis that GC content is higher in regions of more functional contraint was supported. *D. discoideum* and *P. falciparum* are extremely GC poor organisms, where this pressure is likely to be most obvious, however it is worth looking at other species with higher GC contents.

**References**:

Akashi H (2001) Gene expression and molecular evolution. Current Opinion in Genetics & Development 11(6): 660-666.

Chanda I, Pan A, Dutta C (2005) Proteome composition in Plasmodium falciparum: Higher usage of GC-rich nonsynonymous codons in highly expressed genes. Journal of Molecular Evolution 61(4): 513-523.

Chen Y, Emerson JJ, Martin TM (2005) Evolutionary genomics - Codon volatility does not detect selection. Nature 433(7023): E6-E7.

D'Onofrio G, Bernardi G (1992) A universal compositional correlation among codon positions. Gene 110(1): 81-88.

Devreotes P (1989) Dictyostelium discoideum - a model system of cell-cell interactions in development. Science 245(4922): 1054-1058.

Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134(2): 341-352.

Duret L (2006) The GC content of primates and rodents genomes is not at equilibrium: A reply to Antezana. Journal of Molecular Evolution 62(6): 803-806.

Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P et al. (2002) Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. Nucleic Acids Research 30(11): 2515-2523.

Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R et al. (2005) The genome of the social amoeba *Dictyostelium discoideum*. Nature 435(7038): 43-57.

Fryxell KJ, Moon WJ (2005) CpG mutation rates in the human genome are highly dependent on local GC content. Molecular Biology and

Evolution 22(3): 650-658.

Gardner MJ, Hall N, Fung E, White O, Berriman M et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419(6906): 498-511.

Gautier C (2000) Compositional bias in DNA. Current Opinion in Genetics & Development 10(6): 656-661.

Graur D, Li W-H (2000) Fundamentals of Molecular Evolution. Second edition. Fundamentals of Molecular Evolution Second edition: i-xiv, 1-481.

Gutierrez G, Marquez L, Marin A (1996) Preference for guanosine at first codon position in highly expressed Escherichia coli genes. A relationship with translational efficiency. Nucleic Acids Research 24(13): 2525-2527.

Hahn MW, Mezey JG, Begun DJ, Gillespie JH, Kern AD et al. (2005) Evolutionary genomics - Codon bias and selection on single genomes. Nature 433(7023): E5-E6.

Hershberg R, Petrov DA (2010) Evidence That Mutation Is Universally Biased towards AT in Bacteria. Plos Genetics 6(9).

Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of Selection upon Genomic GC-Content in Bacteria. Plos Genetics 6(9).

Jiang RHY, Govers F (2006) Nonneutral GC3 and retroelement codon

mimicry in Phytophthora. Journal of Molecular Evolution 63(4): 458-472.

Kagawa Y, Nojima H, Nukiwa N, Ishizuka M, Nakajima T et al. (1984) High guanine plus cytosine content in the 3rd letter of codons of an extreme thermophile - DNA sequence of the isopropylmalate dehydrogenase of Thermus thermophilus. Journal of Biological Chemistry 259(5): 2956-2960.

Knight RD, Freeland SJ, Landweber LF (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol 2(4): RESEARCH0010.

Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420(6912): 218-223.

Lind PA, Andersson DI (2008) Whole-genome mutational biases in bacteria. Proceedings of the National Academy of Sciences of the United States of America 105(46): 17878-17883.

Lynch M, Sung W, Morris K, Coffey N, Landry CR et al. (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. Proceedings of the National Academy of Sciences of the United States of America 105(27): 9272-9277.

Marais G (2003) Biased gene conversion: implications for genome and

sex evolution. Trends in Genetics 19(6): 330-338.

Morton BR, Bi IV, McMullen MD, Gaut BS (2006) Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. Genetics 172(1): 569-577.

Muramoto T, Suzuki K, Shimizu H, Kohara Y, Kohriki E et al. (2003) Construction of a gamete-enriched gene pool and RNAi-mediated functional analysis in Dictyostelium discoideum. Mechanisms of Development 120(8): 965-975.

Nielsen R, Hubisz MJ (2005) Evolutionary genomics - Detecting selection needs comparative data. Nature 433(7023): E6-E6.

Ohta T (1992) The nearly neutral theory of molecular evolution. Annual Review of Ecology and Systematics 23: 263-286.

Olsen R (2005) How many protein encoding genes does *Dictyostelium discoideum* have? In: F. LW, A. K, editors. *Dictyostelium* genomics. Wymondham, UK: Horizon Bioscience. pp. 265-278.

Ophir R, Itoh T, Graur D, Gojobori T (1999) A simple method for estimating the intensity of purifying selection in protein-coding genes. Molecular Biology and Evolution 16(1): 49-53.

Pascal G, Medigue C, Danchin A (2006) Persistent biases in the amino acid composition of prokaryotic proteins. Bioessays 28: 726-738.

Plotkin JB, Dushoff J, Fraser HB (2004) Detecting selection using a single

genome sequence of M-tuberculosis and P-falciparum. Nature 428(6986): 942-945.

Pozzoli U, Menozzi G, Fumagalli M, Cereda M, Comi GP et al. (2008) Both selective and neutral processes drive GC content evolution in the human genome. Bmc Evolutionary Biology 8.

Rocha EPC, Feil EJ (2010) Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria? Plos Genetics 6(9).

Schwartz S, Meshorer E, Ast G (2009) Chromatin organization marks exon-intron structure. Nature Structural & Molecular Biology 16(9): 990-U117.

Sharp PM, Devine KM (1989) Codon usage and gene-expression level in Dictyostelium discoideum- highly expressed genes do prefer optimal codons. Nucleic Acids Research 17(13): 5029-5039.

Shen JC, Rideout WM, Jones PA (1994) The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Research 22(6): 972-976.

Singer CE, Ames BN (1970) Sunlight ultraviolet and bacterial DNA base ratios. Science 170(3960): 822-&.

Strassmann JE, Zhu Y, Queller DC (2000) Altruism and social cheating in the social amoeba Dictyostelium discoideum. Nature 408(6815):

965-967.

Urushihara H, Morio T, Saito T, Kohara Y, Koriki E et al. (2004) Analyses of cDNAs from growth and slug stages of Dictyostelium discoideum. Nucleic Acids Research 32(5): 1647-1653.